Essays on Inequality and Market Failure

Essays on Inequality and Market Failure

A dissertation presented by

Nathaniel Green Hilger

to

the Department of Economics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Economics

Harvard University
Cambridge, Massachusetts

April 2013

Dissertation Advisor: Professor Raj Chetty Nathaniel Green Hilger

Essays on Inequality and Market Failure

ABSTRACT

This dissertation comprises three chapters. The first chapter develops a research design to estimate the causal effect of parental layoffs and income during adolescence on children's college outcomes, and implements this design on administrative data for the United States. The design compares outcomes of children whose fathers lose jobs before college decisions with outcomes of children whose fathers lose jobs after college decisions. I find that layoffs and unanticipated income losses during adolescence have very small adverse effects on future college outcomes. These effects are smaller than estimates in prior work based on firm closures rather than timing of layoffs. I replicate these larger estimates and show they are driven by selection of workers into closing firms. The findings suggest that relaxing parental liquidity constraints during adolescence will do little to increase enrollment compared to improvements in financial aid, especially for low-income children.

The second chapter, written with my advisor and other colleagues, shows that classroom quality in early childhood has large causal impacts on adult outcomes, and that test score gains can help to identify classroom quality even when these gains fade out over time. We first link administrative data to records from Project STAR, in which 11,571 students in Tennessee and their teachers were randomly assigned to classrooms within their schools from kindergarten to third grade. We then document four sets of experimental impacts. First, students in small

classes are more likely to attend college and exhibit improvements on other outcomes. Second, students who had a more experienced teacher in kindergarten have higher earnings. Third, students who were randomly assigned to higher quality classrooms in grades K-3 -- as measured by classmates' end-of-class test scores -- have higher earnings, college attendance rates, and other outcomes. Finally, the effects of class quality fade out on test scores in later grades but gains in non-cognitive measures persist.

The third chapter explores theoretical properties of markets for "credence goods." Credence goods such as health care involve consumer reliance on expert diagnosis. When consumers observe expert cost functions, competitive markets tend strongly toward efficiency. I argue that consumers do not observe expert cost functions and extend an existing model to incorporate this insight. The key result is that prices and competition no longer eliminate mistreatment.

Contents

ACKNOWLEDGMENTS

I thank my parents Judith C. Green and Christopher A. Hilger for their infinite love and support. I thank my thesis committee Raj Chetty, Andrei Shleifer and Larry Katz for teaching me to think about the world with greater clarity and openness. I also wish to thank Gary Chamberlain, John Pencavel and Danny Yagan. I gratefully acknowledge financial support from the National Science Foundation, the Weatherhead Center for International Affairs, and Harvard University.

I Do Parental Credit Constraints Affect Children's College Choices? Evidence From Timing of Parental Layoffs*

I.A Introduction

Children with top-quartile family income during adolescence are five times more likely to graduate from college than children with bottom-quartile family income during adolescence in the United States (Bailey and Dynarski 2011). These educational gaps have large implications for economic mobility and inequality in future generations. But does family income itself cause higher college attendance, or do these differences stem from correlated factors such as preferences, abilities, beliefs or income at earlier ages? Economic theory suggests that family income can affect children's educational outcomes if non-collateralizability of human capital constrains debt (Schultz 1961, Becker 1994) or if families value education as a consumption good (Lazear 1977, Mulligan 1997). However, the extent to which these mechanisms affect college outcomes remains unclear due to a lack of suitable research designs and data. Some studies argue that family income has substantial causal effects on children's college education (e.g., Acemoglu and Pischke 2001, Belley and Lochner 2007, Oreopoulos, Page and Stevens 2009, Lovenheim 2011), while others find much smaller effects (e.g., Mayer 1997, Blau 1999, Carneiro and Heckman 2002, and Cameron and Taber 2004).

In this paper, I provide new evidence on this question from variation in the *timing* of fathers' layoffs around children's ages of college entry.[1] I implement this research design using selected administrative data on the population of parents and children in the United States and find that the causal effect of unanticipated income on children's college outcomes is positive but small. Furthermore, I replicate the research designs of prior studies in the literature that find much larger effects, and show that these prior results are biased up by selection, reconciling the two sets of

*I am very grateful for the advice and support of my primary advisor, Raj Chetty, throughout this project. I also wish to thank advisors Larry Katz and Andrei Shleifer, as well as John Friedman, Gary Chamberlain, and Danny Yagan for comments that greatly improved the paper. I also thank Josh Angrist, Sam Asher, Thomas Barrios, David Cutler, Nora Dillon, Ed Glaeser, Josh Gottlieb, Jacob Leshno, Joana Naritomi, Arash Nekoi, Amanda Pallais, Jesse Rothstein, Emmanuel Saez, Seth Stephens-Davidowitz, Ian Tomb, and Clara Zverina for helpful comments. The tax data were accessed through contract TIRNO-09-R-00007 with the Statistics of Income (SOI) Division. Research support from the National Science Foundation and the Project on Justice, Welfare, and Economics is gratefully acknowledged.

[1] Time-of-event variation, as opposed to time-of-outcome variation, is widely applied in empirical studies to detect selection problems with cross-sectional estimators (Mayer 1997, Hurst and Lusardi 2004, Rothstein 2009, Coelli 2010, Rege and Votruba 2011) and, less frequently, to identify causal effects (Grogger 1995, Hoynes, Schanzenbach, and Almond 2012).

findings.

My research design compares effects of layoffs that occur before children reach ages of college enrollment with effects of layoffs that occur *after* children reach ages of college enrollment. Layoffs generate large, sudden reductions in family income (Jacobson, Lalonde and Sullivan 1993, Wachter, Song and Manchester 2009, Stephens 2001) that lead to corresponding drops in consumption (Gruber 1997, Stephens 2001, 2004, Chetty and Szeidl 2007). In addition, layoffs may also reduce parental health and happiness (Wachter and Sullivan 2009, Kassenboehmer and Haisken-DeNew 2009), and therefore may have even larger impacts on children than pure income shocks, an issue I revisit below. The research design exploits these sharp, unanticipated changes in resources to identify the causal effect of resources during adolescence on child outcomes.

The key identifying variation is straightforward. Suppose the outcome of interest is college enrollment at age 18. Some children are 17 when their fathers experience layoff. These children make their age-18 enrollment decision in an environment of low parental income, and perhaps lower parental health. Other children are 19 when their fathers experience layoff. These children make their age-18 enrollment decision before layoffs take place, in an environment of high parental income and high parental health. These two groups of children are much more similar to each other than to children who never experience a father's layoff, but their age-18 college decisions take place in very different resource environments. I interpret differences in college outcomes across these two groups of children as the treatment effect of layoff. The research design generalizes the comparison in this example into an event study framework incorporating outcomes and shocks at many ages, and relies on familiar identification assumptions from difference-in-difference (DD) estimation such as parallel trends, no-anticipation and no-manipulation that I confirm empirically.

I implement the research design on the full population of United States taxpayer records. The data set is much larger than typical data sets used to match children and parents in the US, and it contains less measurement error and virtually no attrition. The analysis sample contains millions of children who experience a paternal layoff between 2000 and 2009, and links these children to a rich set of outcomes and family characteristics.

I find that layoffs reduce children's college enrollment by 0.43 percentage points (SE .09) or 1.1% of base enrollment during ages 18-22. Layoffs also reduce enrollment at colleges out of a child's home state (0.529 percentage points or 2.0%), four-year colleges (0.27 percentage points or

1.4%), and non-public colleges (0.1 percentage points or 1.5%), and reduce college quality defined by alumni earnings ($84 or 0.35%). Fathers' layoffs also slightly increase the fraction of children who work during adolescence and college.

The results suggest that 10-15% of the cross-sectional effect of late childhood family income on college enrollment is due to late childhood parental inputs, and that a father's layoff during adolescence likely reduces the net present value of a child's future earnings by about $1,000-3,000. In order to assess the plausibility of these small effects, I examine supplementary data sets on college finance gathered by Sallie Mae and the Department of Education. I find that the effects estimated here are what one would expect from cross-sectional correlations between parental income and spending on college (the Engel curve) in conjunction with the best estimated effects of college price subsidies (Deming and Dynarski 2010). The key insight is that layoffs most likely reduce parent college spending by only $100-500, predicting small effects of layoffs on enrollment even if children respond as strongly to parental contributions as they respond to financial aid. Small income effects are therefore easy to reconcile with existing evidence that children respond strongly to college price, and may even be viewed as weak additional evidence in support of large price effects.

I am also able to measure biases from selection that arise in alternative research designs that do not rely on timing of parental layoffs. Estimates that rely on cross-sectional variation in father's layoffs and control for observable selection into layoff are biased up by 200% in my data.[2] Researchers have tried to address concerns about selection into layoff by restricting to firm closures, since workers displaced by firm closure are not hand-picked by managers for layoff (Oreopoulos, Page and Stevens 2009, Bratberg, Nilson and Vaage 2008, Shea 2000). Surprisingly, I find that firm closures suffer from an even larger selection problem because they do not leave any within-firm "survivors" for use as a control group.

This finding bears on a broad range of empirical applications measuring effects of firm-level shocks on outcomes that are difficult to observe for the same individual before and after shocks take place. The problem can arise for outcomes that are absorbing states such as mortality (Wachter and Sullivan 2009) and disability (Rege, Telle and Votruba 2009). The problem can also arise when

[2] Note my data contain a rich set of observables but lack any measure of children's pre-college academic achievement. Past researchers that control for high school test scores have estimated effects of parental income that are more consistent with those I obtain from timing of layoffs, though with relatively wide confidence intervals due to smaller samples (e.g. Carneiro and Heckman 2002).

shocks and/or outcomes are tied to specific ages, as in the case of estimating childhood shocks on adult earnings (Oreopoulos, Page and Stevens 2009) or college choices (Lovenheim 2011). Results here suggest that it is important to analyze these outcomes in such a way that permits assessment of parallel trends in the outcome itself prior to the shock, rather than relying on other observable characteristics to control for selection across firms. Here I assess the parallel trends assumption by exploiting wide variation in the age at which shocks are experienced, rather than the age at which outcomes are observed. Another approach is to examine firm-level aggregate outcomes rather than individual outcomes (Wachter and Sullivan 2009, Rege, Telle and Votruba 2009). The results here suggest that caution is warranted when firm-level analysis cannot reject problems with selection-on-unobservables across firms.

I am also able to shed some light on mechanisms underlying the main effects by examining heterogeneous impacts across subgroups. Adverse effects of family income shocks on college enrollment are U-shaped in family income levels: smallest at low incomes, larger at middle incomes, and smaller again at high incomes. This non-monotonic pattern is easiest to explain with an income-loss channel. Low-income children rely much less heavily on parents to finance college (Sallie Mae 2011). Counterintuitively, college decisions of the lowest-income children therefore appear least vulnerable to parental income shocks during adolescence. In contrast, smaller effects for high-income children—the second half of the U-shape—could arise if these parents are less liquidity-constrained (Becker 1994), or if college is a smaller share of their total consumption (Mulligan 1997). I also provide evidence that income losses in themselves are sufficient to explain the main results by showing that children reduce enrollment more in families that depend more heavily on fathers' earnings, and in families where fathers are predicted to lose more earnings from layoff. Finally, the evidence suggests that enrollment reductions after layoff stem from *permanent* income declines, not short-term liquidity constraints on parent spending.

The results of this paper contribute to the debate on how transfer programs to low-income households should be structured. Programs that explicitly target children and parents account for $300 billion or 10% of annual federal spending in the U.S.[3] Income support for parents accounts for about half of this; the other half subsidizes or provides child inputs such as education and health

[3] At the federal level, the main income support programs for parents are the Child Tax Credit, the Dependent Exemption, and the Earned Income Tax Credit. The main child input subsidies are education and Medicaid grants for states.

care, bypassing parental preferences. The results here only capture effects of unanticipated income changes during adolescence, holding fixed parental incomes and child inputs at earlier ages, and therefore only bear directly on effects of unanticipated income transfers to parents of older children. However, the findings suggest that reducing input prices may be a much more effective policy to increase child investments than policies that focus on raising parental incomes. For example, in the case of college, the results suggest that input subsidies in the form of financial aid are likely 2-3 *orders of magnitude* more effective at raising college enrollment than policies that relax credit constraints for parents of college-age children. The simplest explanation for small income effects and large price effects is that parents only spend a small share of marginal income on assisting children with college costs.

The paper proceeds as follows. Section II describes the data. Section III describes the empirical strategy. Section IV presents estimated effects of fathers' layoffs on family resources and child outcomes, and performs several robustness checks on the main results. Section V replicates prior work on firm closures, and documents selection-on-unobservables into employment at closing firms. Section VI estimates heterogeneous treatment effects of fathers' layoffs on children by income, wealth, gender, and predicted earnings losses. Section VII discusses the main results in the context of college financing in the U.S. Section VIII concludes.

I.B Data and Summary Statistics

I.B.1 Variables and Sample Restrictions

This paper draws on selected tables from the population of U.S. tax records. Many of the variables used in this paper are described in Chetty, *et al* (2011). The population data sets underlying my analysis have also been used in recent research by Chetty, Friedman and Rockoff (2011), Chetty, Friedman and Saez (2012), and Yagan (2012). The data contain variables on Form 1040 over the years 1996-2009. These variables include adjusted gross income (AGI), marital status, residential location, and various taxes paid and rebates received. I only observe these variables when families file taxes. The data also contain variables on forms filed by institutions on behalf of individuals, or "information returns," for 1999-2009. These variables are observed independent of individual filing. These variables include earnings and deferred compensation (W2), unemployment insurance

(1099G), disability insurance and social security payments (SSA-1099), college enrollment and choice of college (1098T), mortgage interest payments (1098), and interest payments from financial institutions (1099-INT). All of my key results rely exclusively on information returns around the time of layoff. This is important because the probability of filing a tax return varies sharply around layoff, making it difficult to distinguish "real" behavioral responses from patterns induced by changes in filing propensities.

This data set has two large advantages over survey data sets that have typically been used to study links between parents and children in the U.S. First, the main variables contain little recall error, because they are reported by institutions relying on data from actual transactions. Second, the sample size provides a large improvement in statistical power. The two standard intergenerational survey data sets in the U.S., the Panel Study of Income Dynamics and the National Longitudtinal Survey of Youth, both contain a few thousand families, or approximately 1/10,000th (0.01%) of the data used here. These two advantages prove critical to the analysis below. The decrease in measurement error and increase in sample size allow me to employ a non-parametric estimation strategy, estimate small effects with precision, and estimate separate effects on narrowly-defined subgroups.

The data raise three issues that should be discussed, though none of them represent a serious problem for the research design. First, I use unemployment insurance (UI) benefit collection to identify layoffs. UI take-up rates are 72-83% in the U.S. (Currie 2006), and I show below that layoffs defined by sudden UI take-up affect parental earnings and consumption similarly to layoffs studied in prior research (Jacobson, Lalonde and Sullivan 1993, Wachter, Song and Manchester 2009, Chetty and Szeidl 2007).

Second, the 1098T form reflects enrollment of students at Title IV post-secondary institutions in the U.S. This class of institutions covers most, but not all, college students in the U.S. In Appendix 5 I document the extent of imperfect coverage, discuss why it is unlikely to bias my results substantially, and perform two robustness checks using alternative measures of college enrollment. One alternative measure relies on parental claiming of children over age 18 on the 1040, and one measure restricts to colleges that report 1098T's for an extra subset of students. Both measures generate treatment effects consistent with those reported throughout the paper.

The final issue is that I discard 35% of children who I cannot match to fathers through the 1040

form before the children turn 19. I discard 5% of children who are never claimed by parents (some of whom likely immigrated to the U.S. after turning 18), 10% who are only claimed by mothers, and 20% who are claimed by too many adults to identify the father with confidence. Children who I fail to match with fathers tend to have less stable families or families that share claims among adults to increase tax benefits.

These unmatched children have lower college enrollment rates than the children who I match to fathers with confidence. One might worry that I bias my results toward zero by excluding some lower-income children with complex family structures, because these children may be more vulnerable to family income shocks. However, this is unlikely. Appendix 1 documents that the sample of children I link to fathers lines up with the income distributions of comparable male-headed households in the 2001 American Community Survey. The analysis sample therefore still contains a large fraction of low-income children. This allows me to show that the effects of parental layoffs on college enrollment for low-income children in my data are actually smaller than effects on high-income children. As I discuss below, this is most likely due to the fact that low-income children do not rely heavily on parental income to finance college. It is therefore likely that sample restrictions excluding low-income children bias my results *up*, not down. A second reason why the sample restriction is unlikely to account for the main results is that I successfully replicate large cross-sectional effects of firm closures on child outcomes, as found in prior research (Oreopoulos, Page and Stevens 2009).

I now describe the two primary samples used in estimation more precisely: "Layoff" and "Survivor" fathers. These samples contain events from 2000-2009 and outcomes from 1999-2009, along with some 1040-based variables from 1996-1999. I rely on Survivors for removing time-trends from the Layoff group, not for removing selection into the Layoff group.

- **Layoffs.** The layoff sample contains 100% of fathers who experience an event defined as a "layoff." I define a "layoff" as occurring in year T if a father receives positive unemployment insurance (UI) in year T, and receives zero UI in the prior year $T-1$. The no-UI-in-prior-year restriction serves two purposes: it assures that the layoff spell begins in the current year, and it eliminates many repeated short-term layoffs followed by recall, since such layoffs generate UI in consecutive years after the first layoff. I focus on fathers for consistency with prior

research, and because fathers' earnings are a larger share of family income in most families.

- **Survivors.** Survivor fathers are a control group for Layoff fathers. Survivors experience "survival" in year T if they work at a firm that lays off at least one father at T. They also must receive zero UI at $T-1$, to match this restriction on Layoff fathers. By this definition, there are no survivors for workers who lose their job in firm closure, and there are no survivors at firms with no workers in the Layoff sample. Workers are often survivors at some firm in every year they are observed. For example, every worker at a very large firm will be a "survivor" in every year, because very large firms will likely have at least one Layoff father every year. For computational reasons, I therefore take a 30% random sample of survivors. The Survivor sample is propensity-score reweighted to match the Layoff sample on pre-event characteristics.[4] I choose reweighting rather than regression as a method of controlling for these pre-event differences for reasons of convenience.

Note that if a father has N children and suffers K events (layoffs, survivals), he enters the data separately NK times. This means that "parent-based" samples are "adult-based" samples weighted by NK.[5] In the Layoff sample, average N is 2.2 children per father and average K is 1.6 layoffs per father.

I.B.2 Summary Statistics and Cross-Sectional Effects

Table 1.1 displays summary statistics for children at age 18 over 1999-2009. Each sample is described by three columns: one column for children who reach age 18 pre-event, one for children who reach age 18 post-event, and one column containing the percent change from pre-event to post-event. Father's earnings are much lower after layoffs, but are also lower after survival, most likely reflecting

[4] The propensity score is estimated on the fraction of the displacing firm that takes up UI, fixed effects for two-digit NAICS industry of displacing firm, fixed effects for three-digit zipcode of displacing firm, gender, whether the worker has any self-employment income 1996-1999, whether the worker had any deferred compensation in 1999, whether the worker had any mortgage interest payments in 1999, fixed effects for average number of children claimed by worker 1996-1999, fixed effects for age of wife at child's birth (no mother found all coded with same age), fixed effects for age of father in year of layoff, marital status of father 1996-1999 interacted with quartic in total family income 1996-1999, year of layoff interacted with firm size in year prior to layoff interacted with quartic in earnings of father in year prior to layoff. I restrict Layoff and Survivor samples to propensity scores within [-.15, .95] and then reweight the Survivor group to match the Layoff group within each child cohort by event-year cell.

[5] Results reported below are robust to clustering on individual children. I ignore clustering issues within families for computational convenience; this is unlikely to affect precision substantially because the number of families in my sample is very large.

that Survivors work at struggling firms. Child college enrollment is actually higher on average after layoff, but rises by even more after survival. The implied DD for effects of layoff at any time in the past on college enrollment at age 18 is -0.7 percentage points, which is similar to the effects I find using the event study design below. Child earnings fall substantially over time within cohorts due to a strong secular trend that has been documented elsewhere (Aaronson, Park and Sullivan 2007).

Table 1.1: Summary Statistics 1999-2009 for Children at Age 19

Sample:	Layoff			Survivor		
	Pre-Shock	Post-Shock	% Diff	Pre-Shock	Post-Shock	% Diff
Parent Outcomes						
Father's Earnings	51,523	37,699	-26.8%	50,782	47,193	-7.1%
Father Married	0.779	0.754	-3.2%	0.778	0.770	-1.0%
Mother's Earnings	21,440	23,212	8.3%	20,053	21,397	6.7%
Father's UI Benefits	493	1,607	226.0%	230	554	140.2%
Post-Tax Family Income	60,725	53,883	-11.3%	59,480	59,193	-0.5%
Child Outcomes						
Enrollment	0.464	0.498	7.4%	0.469	0.511	9.1%
Years Enrolled 18-22	2.033	2.096	3.1%	2.067	2.138	3.5%
Teen Mother	0.082	0.078	-5.6%	0.0784	0.0721	-8.0%
Earnings	7,359	6,134	-16.6%	7,333	6,124	-16.5%
Freq	3,976,779	3,046,107		17,880,678	15,305,421	

Notes: Survivor sample is propensity-score reweighted to match Layoff sample on observables. "Pre-Shock" includes cells in years before events occur;"Post-Shock" includes cells in years after events occur. "% Diff" is the percent difference between Post-Shock and Pre-Shock columns. Averages pool all cohorts turning 19 during the sample period. All dollar variables deflated to 2009 using CPI-U.

Figure 1.1 plots the cross-sectional effects of mean family income from 1996-1999 on long-term child outcomes in the Survivor sample. Figure 1.1.a plots college enrollment at Age 19 by parental income. The function is approximately linear, with a slope of 0.5 percentage points per $1,000 of family income. This slope combines causal effects of parental assistance with college costs, parental spending on earlier child inputs, and many other factors that correlate with family income such as attitudes toward education, cognitive and noncognitive abilities, and social networks. While the pattern is dramatic, it is interesting to compare this slope to reported effects of financial aid. The literature on financial aid estimates that $1,000 of salient, easily-obtained offered aid increases college enrollment by about 3 percentage points (Deming and Dynarski 2010),[6] or six times the slope in Figure 1.1.a. Also note that $1,000 of offered aid costs much less than $1,000 if only a fraction of children enroll in college, and that financial aid is a temporary transfer. Therefore even if the entire slope depicted in Figure 1.1.a is due to liquidity constraints on parents when children reach college, price effects on enrollment are likely to be much larger than parental income effects.

[6] Decreasing the costs of applying for financial aid also appear to have very large impacts on college enrollment relative to cross-sectional effects of family income (Bettinger *et al* 2009).

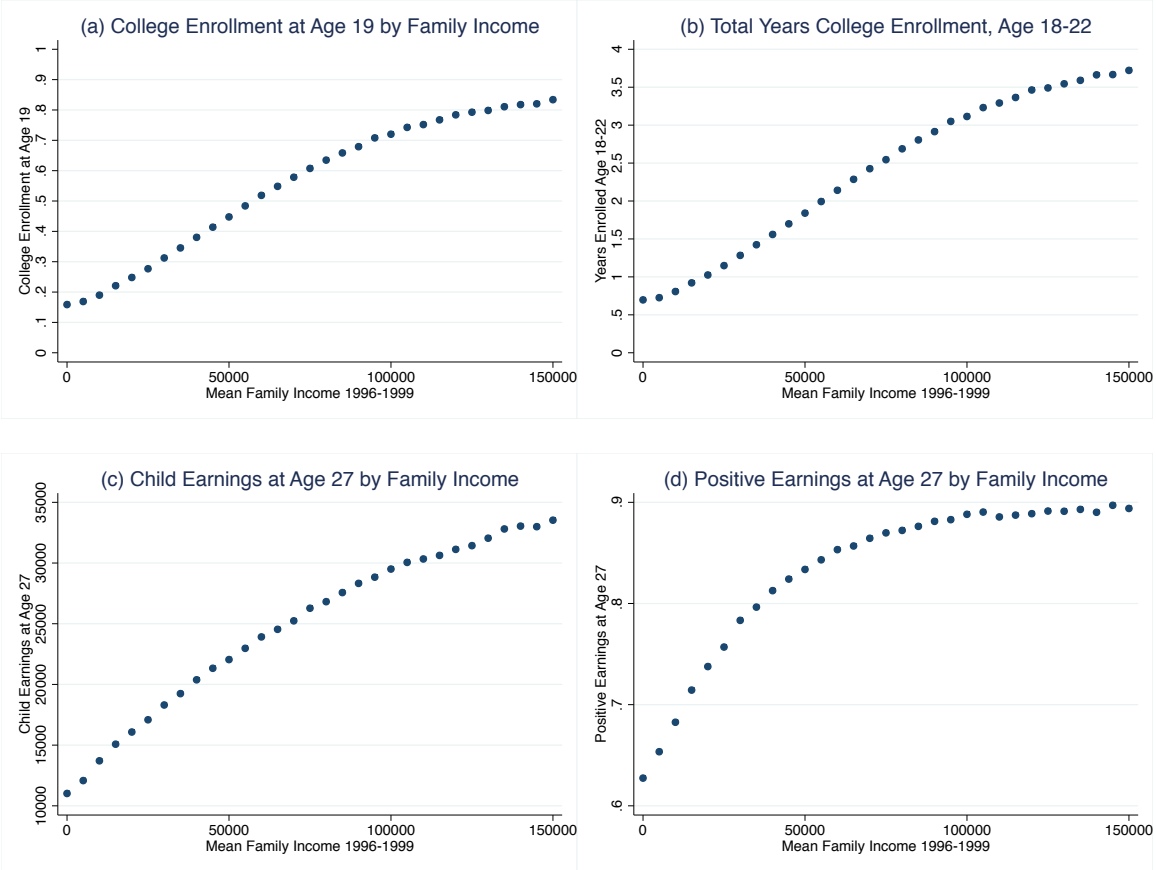


Figure 1.1

Child Outcomes and Parental Income in the Cross-Section: Survivor Sample

Notes: Family income is pre-tax mean family income for claiming father, 1996-1999.

Figure 1.1.b shows the analogous effect of family income on total years enrolled in college over ages 18-22 is .025 years of college per $1,000 of family income. The slopes in Figure 1.1.a and 1.b both understate the difference in child outcomes because they ignore college quality, which also rises rapidly in family income. Figure 1.1.c plots child earnings at age 27, and Figure 1.1.d plots the fraction of children working at age 27, again both by family income. Perhaps in part due to differences in preceding college outcomes, children of richer parents have much higher earnings and are much more likely to work.

These cross-sectional effects of family income serve as benchmarks for interpreting magnitudes of causal effects estimated below. The enrollment effect is consistent with prior work using the National Longitudinal Survey of Youth.[7]

I.C Empirical Strategy

I follow Jacobson, LaLonde and Sullivan (1993) in estimating the dynamic effects of layoffs on parental outcomes using an event-study design, which is a more general form of difference-in-difference (DD) estimation. Let t_O index the year in which an outcome is observed, and t_E index the year in which an event is experienced, or "event-year." Define $k \equiv t_O - t_E$ as "period" or "years after event." Let $g \in \{T, C\}$ index whether a family experiences a "treatment event" T or a "control event" C where the treatment event here is layoff and the control event is survival. Note that both t_O and t_E can provide variation in k. Event studies typically rely on variation in t_O, but for children's college outcomes I will rely heavily on variation in t_E.

Figure 1.2 displays the estimation strategy on data that I generated for illustrative purposes. The figure plots outcomes for a treatment and control group by period. Variation in period may come from t_O, or t_E, or both. These data exhibit a time-invariant selection effect γ into the treatment group, a treatment effect one year after treatment of β_{DD}, and a cross-sectional difference between treatment and control groups one year after layoff of $\beta = \gamma + \beta_{DD}$. This paper improves on prior literature by estimating β_{DD} and γ separately, rather than grouping them together as β.

[7] Author's calculations combining reported statistics in Bailey and Dynarski 2011 and Belley and Lochner 2007.

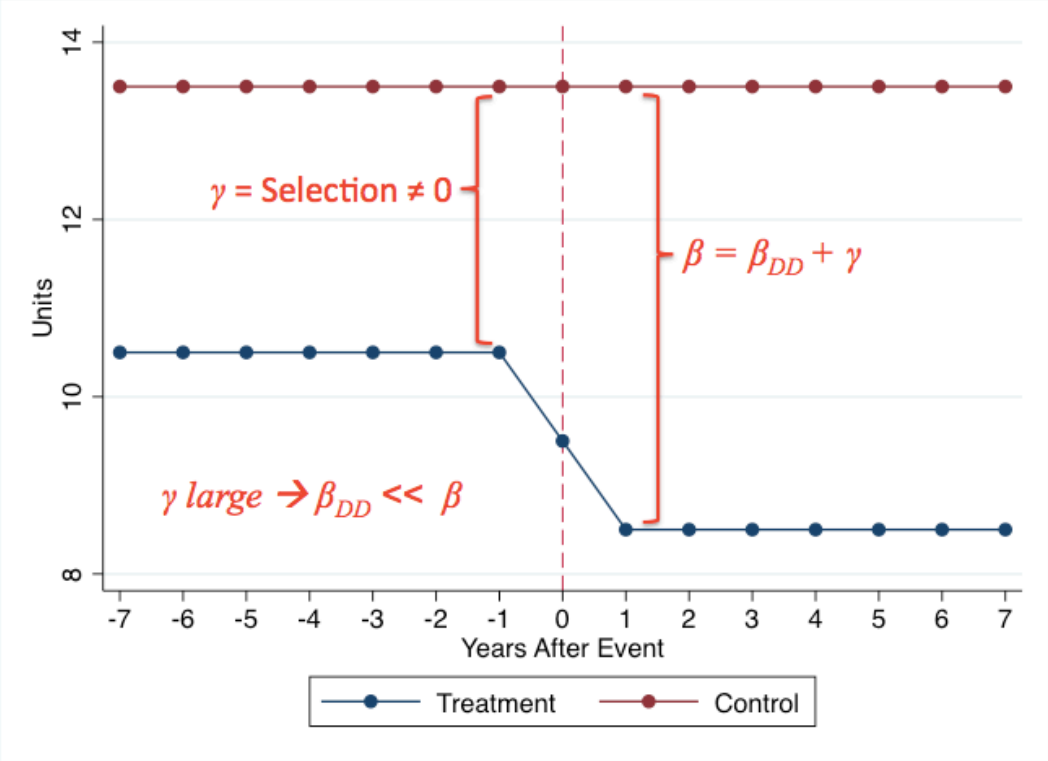


Figure 1.2

Illustration of Event Study for Example Outcome Variable

Notes: This Figure 1.contains no actual data. It plots data that was generated for illustrative purposes.

I now formalize this simple approach. I start with a model for parent outcomes, then slightly modify the approach for estimation of child outcomes. For parents, I estimate an event study model that allows effects of treatment and control events to vary by period:

$$y_{g,t_O,t_E} = \alpha + \sum_{j=k_{\min}}^{k_{\max}} \beta_j^T \cdot I\{g = T, k = j\} + \sum_{j=k_{\min}}^{k_{\max}} \beta_j^C \cdot I\{g = C, k = j\} + \Gamma X_{g,t_O,t_E} + u_{g,t_O,t_E}, \quad (1)$$

$$| \quad k_{\min} < 0 < k_{\max}$$

where α is a constant, β_j^g terms are coefficients on the period dummies, X_{t,t_W} is a vector of observable covariates, Γ is a vector of coefficients, and u_{t,t_W} is an independently-distributed error term. I run this regression on data collapsed into (t_O, t_E) cells, which is the level of treatment variation for parents. Robust standard errors on data grouped at the (t_O, t_E) level with group sizes as sampling weights are identical to those obtained by using micro data and clustering at the (t_O, t_E) level, as long as covariates do not vary within (t_O, t_E) groups, which is true in this application. However, I assign weights of 1 to all cells rather than group size weights because it is critical that each y_{C,t_O,t_E} moment have exactly one corresponding y_{T,t_O,t_E} moment.

For children it is important to allow for an age dimension in the collapsed data, because child's college outcomes vary systematically by age. I also restrict ages to 18-22 for college outcomes, and to ages 14-17 for teen outcomes. This means that, compared to parents, more of the variation in period k comes from variation in event-year t_E than outcome-year t_O, because the age restriction eliminates many values of t_O for each child but places no restrictions on t_E. Let a index age, and write the estimating equation for children as

$$y_{a,g,t_O,t_E} = \alpha + \sum_{j=k_{\min}}^{k_{\max}} \beta_j^T \cdot I\{g = T, k = j\} + \sum_{j=k_{\min}}^{k_{\max}} \beta_j^C \cdot I\{g = C, k = j\} + \Gamma X_{a,g,t_O,t_E} + u_{a,g,t_O,t_E} \quad (2)$$

$$| \quad k_{\min} < 0 < k_{\max}.$$

The only variables I include in X_{a,g,t_O,t_E} are dummies for event-year and cohort where cohort equals $t_O - a$,[8] where now these dummies are interacted with age a. As above, DD estimators are then constructed with linear combinations of the β_k^g coefficients on the period dummies.

In addition to plotting all of the period coefficients in graphs to display the full period trend of treatment effects for several outcomes, I report DD estimators of treatment effects at different time intervals from linear combinations of the period coefficients. I define the "short-run" DD treatment effect as $\beta_{1,-1} \equiv \beta_1^T - \beta_1^C - \left(\beta_{-1}^T - \beta_{-1}^C\right)$, which I also denote in Figure 1.2 and other figures below as β_{DD}, and the "long-run" DD treatment effect as $\beta_{5,-1} \equiv \beta_5^T - \beta_5^C - \left(\beta_{-1}^T - \beta_{-1}^C\right)$. This family of DD estimators β_{k_1,k_2} all assume a standard potential outcome function of the form

$$y_{g,t_O,t_E} = \delta + \gamma \cdot I\left(g = T\right) + \lambda \cdot I\left(k > 0\right) + \beta_{k_1,k_2} \cdot I\left(g = T, k > 0\right) + u_{g,t_O,t_E} \tag{3}$$

$$| \quad k \in \{k_1, k_2\}, \quad k_2 < 0 < k_1.$$

When constructing estimates of DD terms β_{k_1,k_2}, I include full sets of dummies for outcome-years t_O and event-years t_E in the covariate vector $X_{g,t_{O,}t_E}$ in equation (1). These covariates mechanically have no effect on the point estimates of the DD terms β_{k_1,k_2}, but increase precision dramatically.[9]

For the DD terms β_{k_1,k_2} to represent causal effects of treatment, it must be true that

$$I\left(g = T, k > 0\right) \perp u_{g,t_O,t_E}.$$

The economic assumptions required for this orthogonality condition to hold are:

1. Parallel trends in outcomes by k before events: $\beta_j^T - \beta_j^C = \gamma \mid j < 0$.

2. No pre-emptive response at period $k_2 < 0$. Technically, this is implied by the parallel trends assumption, but it can be tested empirically apart from that assumption.

[8] I assume that Layoff and Survivor families have equal event-year and cohort effects. Identification of the β_{k_1,k_2} terms is still possible if we allow separate cohort effects across T and C groups, or separate event-year effects across T and C groups, but not both. Including this richer set of controls does not substantially change the results.

When assumed equal across groups, event-year and cohort effects do not change the point estimates of the $\beta_k^T - \beta_k^C$ or β_{k_1,k_2} terms, though they leave the β_k^g terms themselves unidentified.

[9] Note that these additional controls leave the period dummy coefficients β_k^T and β_k^C unidentified, but preserve identification for the linear combinations of these dummies used to construct DD estimators.

3. No manipulation of k by parents: $k \perp u_{a,g,t_O,t_E}$.

4. No time-varying shocks to families that correlate with layoff: $k \perp u_{a,g,t_O,t_E}$.[10]

I test and/or relax these assumptions below in the event study graphs and in the "Robustness" section.

The key improvement on prior cross-sectional estimates is to allow for arbitrarily-large, time-invariant selection into layoff on determinants of outcomes, i.e. $\gamma \neq 0$ in Equation (3). Prior researchers have estimated the equation

$$y_{a,g,t_O,t_E} = \alpha + \beta \cdot I\left(g = T\right) + \Gamma X_{a,g,t_O,t_E} + e_{a,g,t_O,t_E} \quad | \quad k > 0, \tag{4}$$

and interpreted β as a treatment effect of layoff. If there is selection on unobservables into layoff, however, then $\beta = \beta_{DD} + \gamma \neq \beta_{DD}$.

Note that $\gamma = 0$ can be achieved mechanically by including pre-event outcomes in the propensity-score for outcomes that are roughly similar across ages, such as parental earnings. However, this becomes more difficult for outcomes that are only subject to treatment over a small range of ages, such as 18-22 for college. It is not appealing to include age 18 enrollment in the propensity score, because this requires restricting to shocks that occur at age 19 or later. This would preclude estimating full impacts of layoff on college at ages 18-19, which are the only ages that many children attend college and a key potential complier margin. This would also yield a value for $k_{\min}$ very close to zero, making it hard to assess the parallel trends assumption in the event study graph. The extreme example of this problem occurs when outcomes are only observed at one age. Then individuals for whom the outcome is observed after the treatment event cannot be matched to anyone in the control event group on the basis of pre-event outcomes, because no such outcomes exist. In such cases the only way to control for unobservable selection into treatment is to rely entirely on variation in t_E, since there is no variation in t_O.

[10] Assumptions (3) and (4) have different economic intuitions but are formally identical in this notation.

I.D Impacts of Layoff on Parents and Children

I.D.1 Impacts on Parents

Figure 1.3.a plots $\hat{\beta}_k^T$ and $\hat{\beta}_k^C$ separately for father's earnings around year of layoff. Mean earnings are close in levels and trends prior to layoff, as expected due to inclusion of pre-event earnings in the propensity score. Starting in the year of layoff, earnings of Layoff fathers fall substantially relative to earnings of Survivor fathers. Five years after layoff, recovery is only partial and appears to be slowing down, suggesting permanently lower earnings for Layoff fathers.

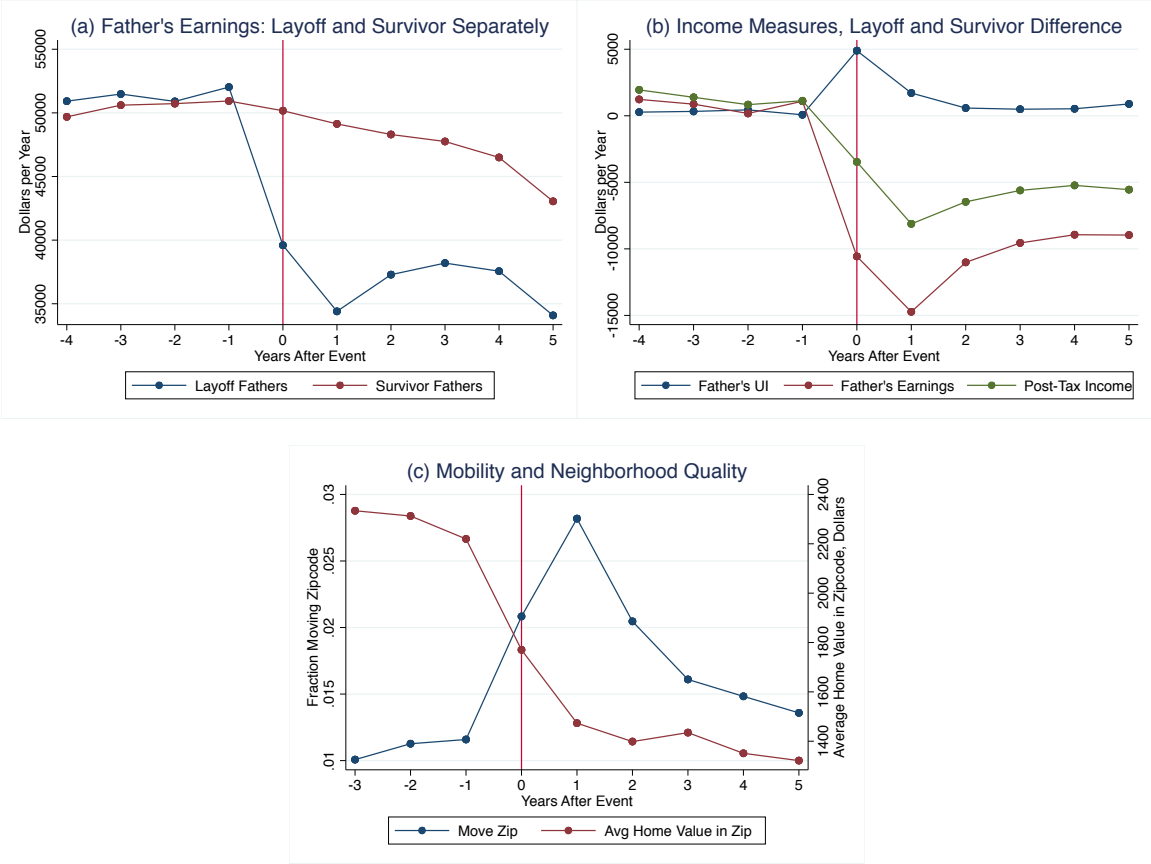


Figure 1.3

Event Studies in Parent Outcomes

Notes: Panel (a) displays simple means of fathers' earnings in the Layoff and Survival groups, after propensity-score reweighting the Survivors. Panel (b) differences these two lines for father's earnings; the lines for father's UI and family post-tax income are constructed analogously. Panel (c) presents analogous differences in means across Layoff and Survivor fathers. Panel (c) loses period -4 because a shift in the X-axis is required to reflect the fact that residency can change prior to tax-filing in April and hence can be affected by layoffs one tax year in the future.

Figure 1.3.b plots the difference between the two lines in Figure 1.3.a, and similar differences for UI and post-tax income. Now each point estimates a cross-sectional difference $\hat{\beta}_k$, and differences between these points estimate DD parameters $\hat{\beta}_{k_1,k_2}$. DD estimates show that layoffs increase UI claims by $\$5,000$ in the year of layoff, but this effect drops off to almost nothing within two years. Average paternal earnings fall by $\$15,000$, after starting at a very comparable level with the reweighted survivor sample, converging to a long-run decline of $10,000. Post-tax family income falls by $10,000 at first, converging to a long-run decline of $6,000. These earnings and income losses are similar to those reported in the prior literature (Wachter, Song and Manchester 2009, Stephens 2001).

Table 1.2 presents short-run and long-run effects of layoffs on many other outcomes of interest, one and five years after layoffs, respectively. Permanent income falls by less than the decline in father's earnings due to insurance from progressive taxation, a $300 increase in DI benefits, and a $500 increase in mother's earnings. The small role of mother's earnings in providing insurance against layoff is consistent with some prior work in the US (e.g., Cullen and Gruber 2000) though results in this literature have been mixed (e.g. Stephens 2001).

Table 1.2: Short-Run and Long-Run Effects of Layoff on Parent Outcomes

	Short-Run (One Year)			Long-Run (Five Years)		
Outcome	Effect	SE	Effect/Base	Effect	SE	Effect/Base
Father's Earnings	-$14,865*	$629	-29.18%	-$9,025*	$538	-17.71%
Post-Tax Family Income	-$8,221*	$490	-13.74%	-$5,641*	$466	-9.43%
Father's UI	$2,427*	$421	674.75%	$555*	$172	154.32%
Mortgage Interest Payments	-$383*	$48	-5.76%	-$460*	$46	-6.91%
Father's DI	$83	$44	45.02%	$297*	$93	161.05%
Mother's Earnings	$271*	$92	1.31%	$506*	$82	2.46%
Median Home Value in Zipcode	-$910*	$374	-0.74%	-$1,487*	$482	-1.20%
Fraction Inter-State Moves (PP)	0.76*	0.08	33.55%	0.01	0.05	0.55%
Fraction Local Moves (PP)	0.52*	0.14	6.78%	-0.05	0.11	-0.61%

Notes: (*) indicates statistical significance at 5% level. Presents DD estimates and standard errors using differences across outcomes for Layoff and Survivor parents, and across periods 1 and -1 (short-run) or periods 5 and -1 (long-run). Effect/Base uses base mean outcomes in years before events occur.

In addition to mother's leisure, the data contain two other measures of consumption: mortgage interest payments and neighborhood quality. Mortgage interest payments reflect spending on owner-occupied housing. Five years after layoffs, families have reduced average spending on owner-occupied housing by $460, or 6.9% of base spending. This number is complicated to interpret because it includes zeros, and because there are treatment effects on switching between ownership and renting. Under the assumption of homogeneous treatment effects on home expenditures across owners and renters, the various opposing biases approximately cancel out and I estimate total spending on housing to fall by about 6-9%.[11] Other research has found that layoffs reduce food spending about one-for-one proportionally with permanent income declines (Stephens 2001). Consumption declines of roughly the same magnitude as permanent income declines are consistent with predictions of a simple lifecycle model (as in Attanasio 1999) at median U.S. wealth/income ratios at middle-age around 2-3 (Bricker *et al* 2012).

A second measure of consumption is neighborhood quality. I find that average home value in residential zipcodes declines by over 1% of base levels five years after layoff.[12] Figure 1.3.c plots average home value against the fraction of families that change residential zipcodes by period. The two measures line up well: layoffs induce sudden increases in mobility and sudden declines in neighborhood quality. Zipcode characteristics capture elements of neighborhood quality for both owners and renters, but are harder to map into spending.

These results paint a clear picture. Total post-tax family income falls sharply following a paternal layoff and only partially recovers over the next five years. Following layoff, families are more likely to move homes, move to less expensive neighborhoods, and reduce home expenditures.

[11] This number is obtained from a simplified model with four groups based on ownership status before and after layoff. Let M refer to owners and R refer to renters, and let 0 refer to periods before layoff and 1 refer to periods after layoff, and let MM, MR, RM, RR refer to ownership status in periods 0 and 1, respectively, so MR refers to families that own homes before layoff and rent homes after layoff. Let Y_0^{MM} equal home expenditures by families in group MM in period 0 and let Y_1^{MM} equal home expenditures by families in group MM in period 1, and similarly define Y_0^{MR}, Y_1^{MR}, and so forth. Let the fraction of the sample accounted for by each group be α_{MM}, α_{MR}, α_{RM}, α_{RR}. Let $\theta = \$460$ be the treatment effect on the full sample. I estimate average spending of families who own homes before layoff to be $10,000, pooling MM and MR groups. Under the assumption that all four groups spend the same amount on housing in period 0 and reduce their expenditures by the same amount θ^*, I can write θ as a function of θ^* and solve for θ^*. The solution is $\theta^* = \frac{\theta - \alpha_{MR} Y_0^{MR} + \alpha_{RM} Y_0^{RM}}{\alpha_{MM} + \alpha_{RM}}$. I estimate $\alpha_{MM} = .675$, $\alpha_{MR} = .025$, $\alpha_{RM} = .025$, $\alpha_{RR} = .275$, and assume $Y_0^{MR} = Y_0^{RM} = \$10,000$. The implied solution is $\hat{\theta}^* = \$657$, or 6.6% of total spending. I obtain a slightly higher estimate of 8.3% from a similar exercise that incorporates the treatment effect on just families who own homes prior to layoff.

[12] I observe similar declines for median home value and fraction of population with college degrees in residential zipcode.

This behavioral response supports a view that layoffs are permanent, unprepared-for,[13] at least partially-uninsured 10% income shocks that reduce total consumption by approximately 10%. If liquidity constraints or consumption commitments operate, then more flexible expenditures may fall by more than 10% in the short-run. The "first stage" of the empirical exercise is therefore powerful: layoffs reduce family income and consumption.

I.D.2 Impacts on Children

I now turn to the main results of the paper: the effects of fathers' layoffs on child college outcomes. Figure 1.4.a plots child college enrollment in Layoff and Survivor groups by period as estimated in Equation (2).[14] This is very similar to Figure 1.3.a, but with father's earnings replaced by child college enrollment, and with child age restricted to 18-22.[15] The selection effect γ accounts for a large share of the cross-sectional difference β. This is because college enrollment before events cannot be included in the pre-event propensity score reweighting of Survivor families to Layoff families, and because the treatment effect on children is small. The treatment effect can be seen more clearly in Figure 1.4.b, which differences the two lines in Figure 1.4.a, and is analogous to the line for father's earnings in Figure 1.3.b. Here we see that $\beta_{1,-1}$ is estimated precisely at 0.43 percentage points (t-statistic over 4), whereas the estimated cross-sectional effect $\hat{\beta}$ is nearly three times this large. This causal effect of layoff is 8-14% of the the prediction that would be obtained from the cross-sectional effect of income on enrollment in Figure 1.1.a, depending on whether we use short-run or long-run income losses from layoff in the prediction.

[13] I say "unprepared-for" instead of "unanticipated" to reflect the results in Stephens (2003), who finds that workers can partially predict layoff but do not appear to use these predictions in their consumption plans.

[14] Recall that the period effects themselves are not identified, but the difference between period effects in Layoff and Survivor groups are identified. This means that the slopes of the lines in Figure 3.a are not interpretable, but differences between the two lines are interpretable.

[15] Effects are also detectable on individual ages 18, 19, ..., 22, with negative and very significant effects at all ages except for 21, and the largest effect at age 19.

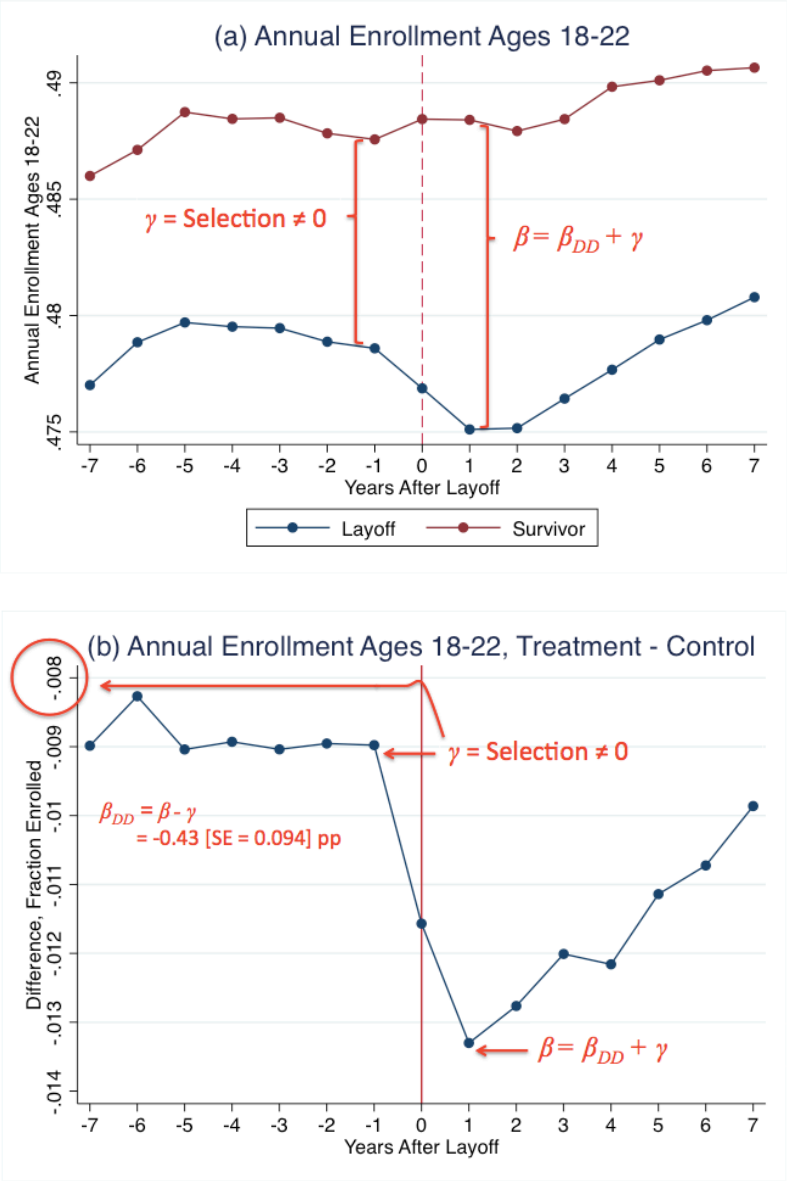


Figure 1.4

Event Study in Child College Enrollment, Pooling Ages 18-22

Notes: Panel (a) displays the period*layoff and period*survival coefficients for periods -7 to 7 from Equation (3) in the text, estimated with cohort*age and event-year*age fixed effects and no constant term. Panel (b) displays the difference between these coefficients, by period *k*.

Figure 1.4.b exhibits several attractive features. The flat line for $k < 0$ indicates that the parallel trends assumption underlying DD estimation is reasonable. Layoffs that occur in the year of a college decision where $k = 0$ reduce enrollment less than layoffs that occur one year before the decision where $k = 1$, most likely because at $k = 0$ some children have already paid for college and started taking classes by the time their fathers are laid off. This pattern suggests that college decisions do not anticipate layoffs even a few months in advance. Finally, the upward slope over $k > 0$ suggests that layoffs have larger short-run effects on enrollment than long-run effects.

Table 1.3 calculates treatment effects during ages 18-22 on a variety of outcomes, along with t-statistics and base levels. The treatment effect on college enrollment is 1% of base enrollment. Children experiencing a father's layoff are also less likely to attend college out of state and less likely to attend a four-year university. These results suggest that children attend lower-priced colleges, in addition to reducing enrollment.

Table 1.3: Effects of Paternal Layoff on College-Age Children: Ages 18-22

Outcome	Effect	SE	T-Statistic	Base	Effect/Base
Percentage Points					
College Enrollment	-0.432*	0.094	-4.580	40.66%	-1.06%
College Enrollment: Out of State	-0.529*	0.115	-4.610	25.90%	-2.04%
College Enrollment: Four-Year	-0.27*	0.114	-2.362	19.61%	-1.38%
College Enrollment: Non-Public	-0.101*	0.047	-2.174	6.79%	-1.49%
College Quality: > $20,000	-0.407*	0.088	-4.620	38.39%	-1.06%
College Quality: > $30,000	-0.286*	0.077	-3.700	28.73%	-1.00%
College Quality: > $40,000	-0.092*	0.043	-2.137	9.15%	-1.01%
College Quality: > $50,000	-0.024	0.020	-1.215	1.84%	-1.30%
College Quality: > $60,000	-0.008	0.008	-1.034	0.48%	-1.75%
Earnings > 0	0.181*	0.064	2.805	84.97%	0.21%
Earnings > $2,000	0.209*	0.074	2.814	71.89%	0.29%
Earnings > $10,000	0.142	0.074	1.910	32.47%	0.44%
Dollars					
Earnings	$0.77	$17.77	0.043	$8,488	0.01%
College Quality: Alumni Earnings	-$83.61*	$19.86	-4.210	$23,930	-0.35%
Family Income	-$8,138.09*	$219.53	-37.071	$59,832	-13.60%

Notes: (*) indicates significance at 5% level. Presents DD estimates and standard errors using differences across outcomes for children of Layoff and Survivor parents, and across periods 1 and -1. Effect/Base uses base mean outcomes before events occur.

These choices translate into lower-quality colleges based on a measure of college quality developed in Chetty *et al* (2011). The measure calculates mean earnings at age 29 for individuals enrolled in each college at age 20 in 1999. Chetty *et al* (2011) shows that this measure is highly correlated with U.S. News and World Report college rankings for the top 125 colleges that the magazine ranks. The alumni-earnings quality index has the advantage of covering all Title IV post-secondary schools in the U.S. Children who are not in college at age 20 are treated as if they are in their own college and assigned age-29 earnings like all other colleges, which for them is about \$16,000. For comparison, earnings of individuals in a top college at age 20 are about \$65,000 at age 29.

By this measure, fathers' layoffs reduce the quality of college that their children attend by \$84 or 0.35%. To see effects across the distribution of college quality I calculate treatment effects on dummies for enrollment in colleges above various quality cutoffs, e.g., \$20,000, \$30,000, and so forth up to \$60,000. The fraction of children above the lowest quality cutoff falls by almost the same amount as the fall in total enrollment, consistent with most compliers on the enrollment margin attending low-quality colleges. However, there are also substantial declines in the fraction of children attending higher-quality colleges. This is consistent with reductions in college enrollment at out-of-state and four-year colleges, and with children responding along a quality margin as well as an enrollment margin.

Children also adjust earnings during college-going ages. Once again, to distinguish extensive and intensive margin responses I calculate treatment effects on dummies for earnings above three cutoffs: \$0, \$2,000, and \$10,000. Results suggest that layoffs push more children to get "real jobs" that yield substantial earnings during college. These effects are all under half of one percent of base earning and earning-cutoff levels. These effects could be driven mechanically by shifts out of college enrollment and into work, rather than higher labor supply while in college. To distinguish these two explanations I also estimate Equation (2) for outcome variables that interact earnings cutoffs with college enrollment. I find no effect on these variables, but this is not conclusive because the children who stop enrolling may have had higher base earnings than other children, masking an increase in earnings among children who remain enrolled.

The approach can estimate treatment effects on outcomes during adolescence by restricting to outcome-ages $a \in [14, 17]$ rather than $a \in [18, 22]$. The key outcomes I examine are earnings and

teen pregnancy for girls. There is a significant response on the extensive earnings margin: parental layoffs increase the likelihood that children will earn positive formal sector earnings by 1.4%, and raise the probability that children earn at least $2,000 by 1.7%. There are no significant effects on average earnings in levels or logs. There are no significant effects on teen birth rates for girls, although these results are underpowered. While there is substantial evidence that higher family income reduces child labor supply for low-income families in developing countries (Edmonds 2007), to my knowledge these results are the first quasi-experimental evidence for this relationship in a developed country (XX cite panel data studies on child labor laws in US, and scale back this claim, fixed effect models are similar to my approach).

The results suggest that children adjust to family resource shocks along a variety margins. Some children forego college, some enroll closer to home or at two-year rather than four-year colleges. Some children enroll at lower-quality colleges that may be cheaper for other reasons. Some children may increase earnings to remain in college, though this is hard to observe empirically. There are several other potential adjustment margins that I do not observe. One such margin is informal earnings. Anecdotally, many college students work as waiters, bussers, bartenders, babysitters, and tutors.[16] In these occupations, consumers often pay workers directly through tips or cash wages, and W2s may fail to record a higher fraction of compensation. A second potential margin is consumption. The average college student consumes about $10,000 of goods on top of tuition and fees for college, including housing, food, transportation, entertainment, clothing, and vacation (Paulin 2001). A third potential margin I do not observe is changes in parent or student borrowing and financial aid. If college enrollment and college quality are valuable investments, many children may choose to adjust these many other margins instead, as in Leslie (1984) and Keane and Wolpin (2001).

I.D.3 Robustness

I now provide evidence in support of the main identifying assumptions, as well as estimators that rely on weaker versions of these assumptions. I present the results by column in Table 1.4.

[16] See, for example, <http://www.onlinecertificateprograms.org/blog/2010/20-most-common-jobs-for-college-students/> and <http://www.kiplinger.com/columns/starting/archive/great-part-time-jobs-for-college-students.html>, accessed on 10/15/12.

Table 1.4: Treatment Effects Under Alternative Assumptions

Outcome (Percentage Points)	1	2	3	4	5
College Enrollment	-0.442	-0.418*	-0.426*	-0.341*	-0.313*
	(0.479)	(0.106)	(0.093)	(0.088)	(0.155)
College Enrollment: Out of State	-0.623	-0.482*	-0.604*	-0.391*	-0.267
	(0.701)	(0.114)	(0.103)	(0.085)	(0.178)
College Enrollment: Four-Year	-0.253	-0.274*	-0.387*	-0.381*	-0.104
	(0.724)	(0.113)	(0.1)	(0.078)	(0.184)
College Enrollment: Non-Public	-0.087	-0.078	-0.171*	-0.135*	-0.125
	(0.329)	(0.048)	(0.045)	(0.039)	(0.082)
Earnings > 0	0.014	0.163*	-0.058	0.084	0.097
	(0.295)	(0.072)	(0.066)	(0.057)	(0.119)

Notes: (*) indicates significance at 5% level. Standard errors in parentheses. Column (1) displays treatment effects using only Layoff sample, as described in Appendix 3. Column (2) displays treatment effects relaxing the parallel trends assumption to a linear differential trends assumption, as described in Appendix 4. Column (3) displays treatment effects that difference across periods -3 and +1. Column (4) presents treatment effects that difference across event-ages before 18 and event-ages after 22, allowing for intertemporal substitution of outcomes within ages 18-22 and allowing for decisions to be made only one time rather than independently year-by-year. Column (5) displays treatment effects using mass layoffs rather than all layoffs.

I employ Survivors throughout my analysis in order to increase precision. However it is possible to estimate treatment effects using only Layoffs. In Appendix 3 I derive an estimator that uses only the Layoff sample. Column (1) of Table 1.4 displays treatment effects using this approach. The estimator yields results that are consistent with those reported above, but are much less precise. The loss of precision occurs because Survivors non-parametrically control for cohort by event-year shocks in the Layoff sample. Cohort shocks occur because of nonlinear secular trends in college outcomes. Event-year shocks occur because selection into layoff on children's college outcomes (and most likely other measures of family achievement) is counter-cyclical: firms only lay off their least productive workers during booms, but lay off higher-productivity workers during recessions.[17] Interactions between cohort and event-year trends are harder to interpret, but turn out to be important relative to the size of treatment effects.

The β_{k_1,k_2} estimators assume parallel trends in outcomes prior to events. When estimating $\beta_{k_1,-1}$ for some k_1, as I do above, I require parallel trends in outcomes with respect to k for $k < 0$. There is no evidence to reject this assumption for college enrollment on the full sample in Figure 1.4.b. However, to check this on outcomes other than college enrollment in Table 1.3, I also estimate treatment effects that allow for linear differential trends in outcomes with respect to k for $k < 0$. This would arise if the line for $k < 0$ in Figure 1.4.b were not flat, but rather had some non-zero, linear slope. This is a weaker version of the parallel-trends assumption, and can be viewed as a triple-difference estimator. In Appendix 4 I derive formulas for the point estimates and standard errors of such an estimator. Results are presented in Column (2) of Table 1.4. These estimates are nearly identical, and remain precise.

The β_{k_1,k_2} estimator assumes that families do not reduce spending on college in anticipation of layoff $-k_2$ years ahead of time. I have used $k_2 = -1$, assuming that families do not pre-emptively reduce spending one year ahead of time. It is possible that families anticipate and respond to layoffs one year before they occur. I therefore estimate treatment effects $\beta_{1,-3}$, requiring that families not reduce spending pre-emptively three years before layoffs take place. The results are presented in Column (3) of Table 1.4. The results on college are nearly identical, while the effect on fraction of children working is no longer significant. There are also substantive reasons to believe that families do not smooth college contributions in anticipation of layoff. First, families adjust

[17] Mueller (2012) finds a similar pattern for father's pre-layoff earnings.

spending on housing only after layoffs occur, and the size of the adjustment is similar to the decline in permanent income. This is consistent with effects of layoffs on food expenditures in the PSID (Stephens 2001). Moreover, evidence in Stephens (2004) suggests that families do not incorporate their (limited) idiosyncratic knowledge of future layoff propensities into their spending plans.

The β_{k_1,k_2} estimator assumes that children choose outcomes independently at each age 18-22. This would be violated if, for example, starting college involved a fixed cost, so that marginal costs of continuing after one's first year are relatively low. The opposite extreme assumption is that children make college enrollment decisions for all ages 18-22 at a single point in time, say age 17 or 18. To address this I average outcomes over ages 18-22 and compare this value for children experiencing events before age 18 with children experiencing events after age 22, continuing to use the DD approach above. All of the variation in k now comes from event-time t_E. Column (4) of Table 1.4 presents the results. The results are noisier but similar. This also suggests that intertemporal substitution of outcomes within the age 18-22 age window, such as delaying college for a year until one's parents recover from layoff, does not account for the treatment effects.

I assume that parents do not carefully postpone being laid-off until their children have enrolled in college, and that layoffs are not driven by time-varying shocks to other parental resources such as illness or divorce. I test this by examining results of mass layoffs. I define mass layoffs for firms that employ over 30 workers in the period $k = -1$ just before layoff, and in which at least 20% of workers claim UI in period $k = 0$. This is not a perfect test, but mass layoffs are driven somewhat less by idiosyncratic factors than the average layoff. Column (5) of Table 1.4 shows that effects of mass layoffs on children are similar to layoffs in the full sample, though once again considerably noisier. The same result will be shown to hold below for firm closures.

I match Survivors to Layoffs using a propensity-score reweighting procedure. I can also estimate the period effects in Equation (2) using regression to control for observables, rather than propensity-score reweighting. I do this only for college, because each outcome requires a separate regression. This approach requires me to estimate coefficients on cohort by event-year dummy variables, then use these coefficients to estimate Equation (2). The estimated treatment effect on annual enrollment during ages 18-22 using this approach is almost identical to the results above: 0.43 percentage points (SE .110).

I.E Impacts of Firm Closures

Recent work on parental layoffs and child outcomes has attempted to address the potential endogeneity of parental layoff by examining job loss due to firm closure (Oreopoulos, Page and Stevens 2009, Rege, Telle and Votruba 2011, Bratberg, Nilson and Vaage 2008). This approach eliminates selection into layoff within firms, but does not eliminate selection into closing firms. There are many reasons why lower-quality workers may wind up at less-profitable firms, and there exists some evidence that firms heading for closure wind up employing less productive workers (Brown and Matsa 2012, Abowd, Kramarz and Margolis 1999). The event study approach developed above allows me to test directly for selection into employment at closing firms.

I first define two additional samples that are analogous to the Layoff and Survivor samples above.

- **Closures.** Fathers experience "closure" in year T if they work at a firm in year T that never issues another W2 to any worker after year T. Note I observe firm closures, not plant closures. I impose the same restrictions described in Oreopoulos, Page and Stevens (2009) in order to compare my results with theirs on fathers and sons in Canadian administrative tax records. The sample restrictions are described in more detail in Appendix 2.

- **Non-Closures.** Non-Closure fathers are a control group for Closure fathers. Non-Closure fathers are a 30% random sample of the U.S. population of fathers. Non-Closure fathers experience a placebo event at T generated by a random variable, imposing the same restrictions as in the Closures sample. As with the Survivors, I propensity-score reweight Non-Closure fathers to match the Closure sample on pre-event characteristics.[18]

Summary statistics for these samples are presented in Appendix 2, and are similar to those for the Layoff and Survivor samples.

Figure 1.5.a plots earnings of Closure and Non-Closure fathers by period. This figure is analogous to Figure 1.3.a above, but treatment and control events are now Closure and Non-Closure rather than Layoff and Survival. Note that the Closure sample is less than 1% of the size of the

[18] I include fewer variables in the Non-Closure propensity score than in the Survivor propensity score. The Non-Closure propensity score includes dummies for event-year, marital status, interacted discretized earnings in the two years prior to closure in $10,000-width bins, dummy for positive 401(k) compensation in 1999, dummy for Schedule C profit in 1999, dummy for mother's age at child's birth under 22, and sex of child.

Layoff sample. Despite the smaller sample, the figure displays the same basic pattern as in Figure 1.3.a. Fathers' earnings in the two groups are close and track each other before events take place, due to the inclusion of father's pre-event earnings in the propensity score. Firm closures have somewhat larger negative short-run impacts $\hat{\beta}_{1,-1}$ than layoffs on father's earnings. The associated short-run post-tax family income loss is about $8,000.

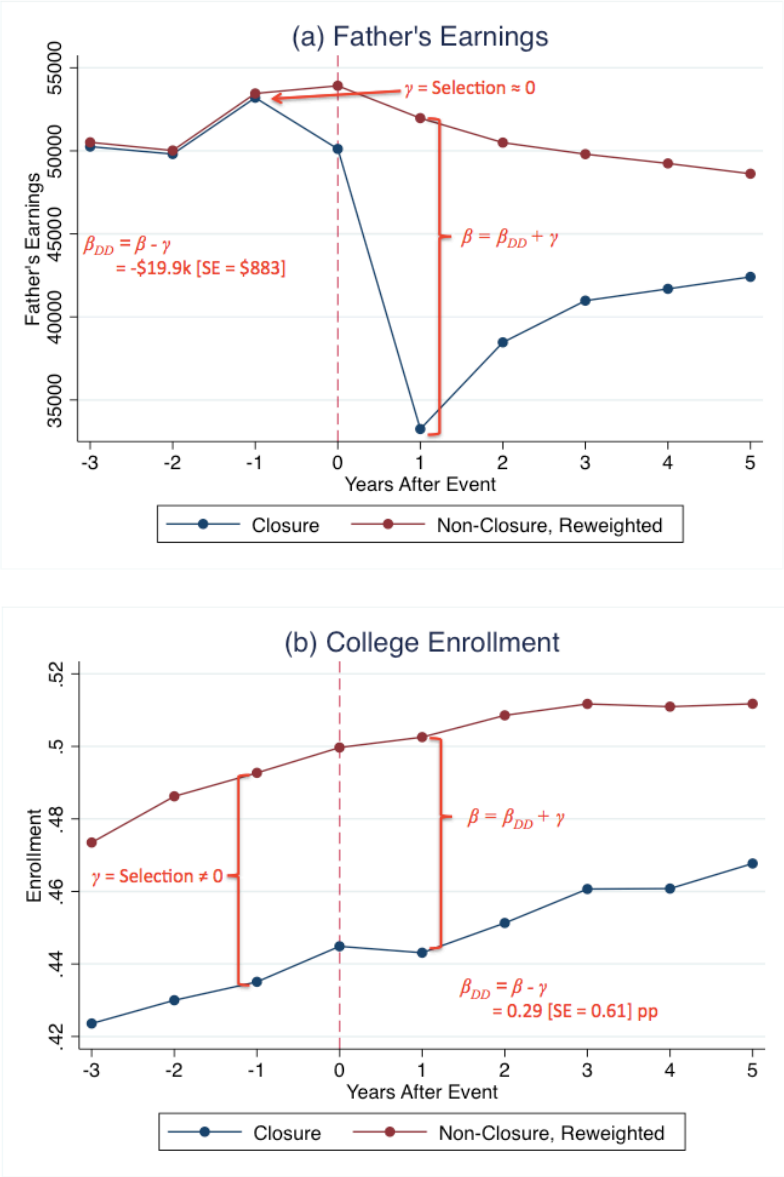


Figure 1.5

Event-Time Studies of Firm Closure

Notes: Panel (a) displays the period*closure and period*non-closure coefficients for periods -3 to 5 from Equation (1) in the text, estimated on father's earnings with outcome-year and event-year fixed effects and no constant term. Panel (b) displays the period*closure and period*non-closure coefficients for periods -3 to 5 from Equation (3) in the text, estimated on child's enrollment with cohort*age and event-year*age fixed effects and no constant term.

Figure 1.5.b plots child's college enrollment by period, as in Figure 1.4.a. The cross-sectional estimate $\hat{\beta}$ is now 6 full percentage points. This is similar to the result obtained by Oreopoulos, Page and Stevens (2009) on a child's earnings: layoffs appear to have larger effects than would be predicted from income losses, even if the entire cross-sectional effect of income in Figure 1.1.a were causal. However, the event study shown in Figure 1.5.b suggests that virtually all of this effect is due to the selection term γ. For this reason, the implied short-run DD estimate $\hat{\beta}_{1,-1}$ is indistinguishable from both zero and from the estimate of $\hat{\beta}_{1,-1}$ obtained using the much larger Layoff sample above.[19]

Oreopoulos, Page and Stevens (2009) also find that closures appear to have larger effects on children in lower-income families. I can test whether this pattern could be generated spuriously by differential selection into firm closure by family income quartiles. Figure 1.6 indeed finds this pattern for college enrollment. The cross-sectional estimated treatment effect $\hat{\beta}$ estimated separately at each family income quartile is twice as large in the bottom quartile as in the top quartile, and this difference is statistically significant. In contrast, none of the DD estimates $\hat{\beta}_{DD}$ are statistically different from zero, from each other, or from the main 0.43 percentage-point effect estimated above, at the 5% level. Below I estimate DD treatment effects across the income distribution more precisely using the Layoff sample.[20]

[19] The standard error on the DD estimated impact of firm closure is approximately ten times larger than the standard error on the DD estimated impact of layoff, because the sample is approximately 100 times smaller for closure than for layoff, and standard errors fall with the square of the sample size.

[20] It should be noted that there are some important differences between this study and Oreopoulos, Page and Stevens (2009). These authors study child earnings in young adulthood, whereas I study college outcomes. I am able to show that layoffs at ages 16-17 reduce earnings at age 25 by at most 1%, but due to my short panel I cannot study effects of earlier shocks on later earnings. Effects on later earnings could be larger than effects on college enrollment would lead us to expect. This appears to be true for teacher effects at younger ages in Chetty, Friedman and Rockoff (2012): if college raises future earnings by 10%, then college only mediates at most 10-20% of the effect of teacher quality on students' future earnings. It is not known how these mechanisms compare for shocks to parental resources during adolescence. Second, these authors study firm closure in Canada, whereas I study firm closure in the U.S. Third, the Closure and Non-Closure samples studied by these authors appear more similar to each other on fathers' and childrens' outcomes prior to controlling for observable characteristics than the analogous samples in my data. In my data Closure and Non-Closure fathers have significantly different earnings, family structures, and other covariates prior to propensity-score reweighting. This difference may stem from larger economic and social heterogeneity in the U.S. relative to Canada. Therefore it is possible that these explanations, rather than selection into firm closure, drive the difference between the small estimated effects on college in this paper and the large estimated effects on child's future earnings in their paper.

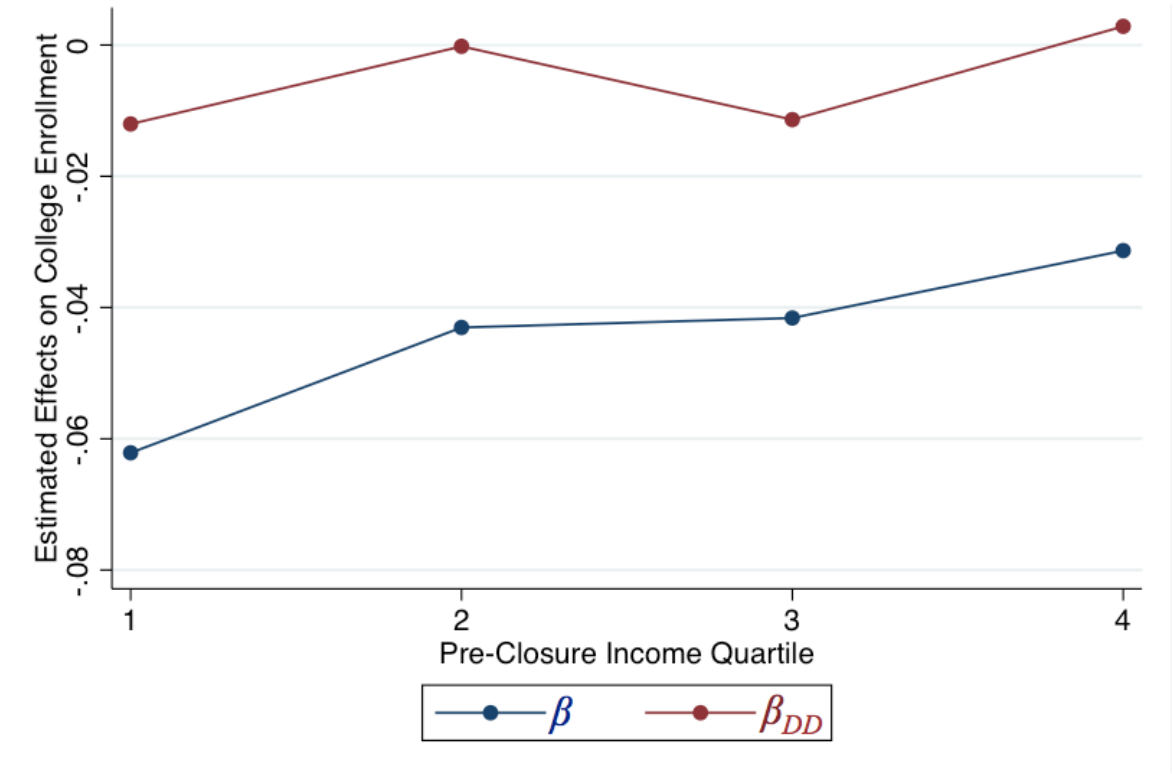


Figure 1.6

Estimated Effects of Firm Closure on Enrollment Ages 18-22 by Income Quartile

Notes: The line labeled $\boldsymbol{\beta}$ displays estimated treatment effects on annual enrollment during ages 18-22 from cross-sectional equation (2). The line labeled $\boldsymbol{\beta_{DD}}$ displays estimated treatment effects on annual enrollment ages 18-22 from the DD equation (3). Both estimates control for observable variables using the same propensity-score reweighting of non-closures to match closures.

These results suggest that allocations of workers across firms are highly endogenous to some types of hard-to-observe worker quality, and that children's propensity to enroll in college inherits these worker quality differentials. In this case the use of firm closures rather than layoffs actually *increases* the selection component γ of the cross-sectional estimate β, because Non-Closures are a worse control group for Closures than Survivors are for Layoffs.

The findings here may pose problems for studies that cannot easily control for selection on the dependent variable into firms that close, initiate mass layoffs, or exhibit other strong signs of failure or success. The standard ways to control for selection on a dependent variable are DD, fixed effect, or event study designs. Selection presents a larger challenge when outcomes cannot be observed both before and after shocks take place for any particular individual. This problem afflicts firm-level shocks on outcomes that are absorbing states such as mortality (Wachter and Sullivan 2009) and disability (Rege, Telle and Votruba 2009). The problem also afflicts studies of firm-level shocks on workers' children where shocks occur long before outcomes as in Oreopoulos, Page and Stevens (2009) and Atkin (2009), or where child outcomes only occur at very few ages such as college outcomes in Lovenheim (2011) and Lovenheim and Reynolds (2012). It is striking that many of these studies estimate very large impacts compared to the relevant cross-sectional benchmarks.[21]

These considerations suggest using alternative methods to establish parallel trends in outcomes before events. Age-specific outcomes such as college enrollment can often be observed before shocks for some individuals and after shocks for other individuals. In such cases the researcher can establish parallel trends in event-time t_E around the age of outcome, which is roughly the approach I take here[22]. Outcomes that absorb workers out of employment require linking workers to firms several periods prior to the shock, i.e. tenure restrictions, which are standard in much of the firm closure literature. Parallel trends can then be established in period-level aggregate outcomes over outcome-

[21] Sullivan and Wachter (2009) find that displacements–some of which are voluntary–increase mortality rates by at least as much as would be predicted from the entire cross-sectional effect of income on mortality, and suggest that layoffs affect health through mechanisms other than income. Atkin (2009) finds effects of maternal employment on child height that are much too large to be explained by maternal earnings, and suggests many explanations why stable maternal employment might benefit children. Lovenheim (2011) and Lovenheim and Reynolds (2012) find effects of home price appreciation on college outcomes that are an order of magnitude larger per dollar of marginal wealth than those estimated here from layoffs, despite the fact that the negative wealth shocks studied here should have larger impacts if credit constraints operate.

[22] If I restrict Equation 2 to outcomes at a single age, such as college at 18, this is exactly what I do. In practice I pool ages 18-22 to maximize power, thereby also exploiting a small amount of variation over outcome-time t_O within individuals between ages 18-22. The results are similar either way.

time t_O around the time of the event, for treatment and control groups. In all cases, If a study exploiting variation in firm performance lacks sufficient statistical power to distinguish effects due to selection-on-unobservables (γ) in periods of non-exposure to treatment ($k < 0$) from total effects ($\beta_{DD} + \gamma$) in periods of exposure to treatment ($k > 0$), then results should be interpreted with caution, especially when they are surprisingly large.

I.F Heterogeneity and Mechanisms

I now explore heterogeneity in treatment effects. I first compare effects for males and females. I then compare effects among various subgroups to understand the mechanisms driving the treatment effects. The key mechanisms to distinguish are (1) income effects versus non-income effects such as family stress or turmoil, and (2) investment effects driven by liquidity constraints versus consumption effects driven by permanent income.

I.F.1 Child Gender

Table 1.5.a examines effects on males and females separately. Results are slightly larger for females across most outcomes, but none of the differences in college outcomes are statistically-significant. These differences are larger when estimating treatment effects that allow for linear differential trends in outcomes, rather than imposing a constant selection effect γ, as described in Appendix 4. Larger effects for girls echo findings in several other studies of monetary incentives for academic achievement and college enrollment, summarized in Angrist and Lavy (2008, p. 25-27). The reasons for this disparity are not well-understood.s

Table 1.5: Effects of Paternal Layoff on College-Age Children: Ages 18-22

(a) Gender

	Male			Female		
Outcome	Effect	SE	Effect/Base (%)	Effect	SE	Effect/Base (%)
Percentage Points						
College Enrollment	-0.331*	0.100	-0.93%	-0.55*	0.118	-1.19%
College Enrollment: Out of State	-0.463*	0.114	-2.08%	-0.611*	0.137	-2.04%
College Enrollment: Four-Year	-0.209	0.111	-1.25%	-0.342*	0.138	-1.50%
College Enrollment: Non-Public	-0.04	0.052	-0.70%	-0.17*	0.061	-2.15%
College Quality: > $20,000	-0.316*	0.095	-0.94%	-0.512*	0.109	-1.18%
College Quality: > $30,000	-0.25*	0.085	-0.98%	-0.33*	0.100	-1.02%
College Quality: > $40,000	-0.031	0.053	-0.37%	-0.159*	0.059	-1.59%
Earnings > 0	0.113	0.082	0.13%	0.255*	0.076	0.30%
Earnings > $10,000	0.054	0.101	0.15%	0.245*	0.089	0.84%
Dollars						
Earnings	-$39.36	$26.63	-0.42%	$47.19*	$16.76	0.62%
College Quality: Alumni Earnings	-$59.01*	$21.65	-0.26%	-$112.05*	$24.95	-0.45%
Family Income	-$8,052.56*	$224.12	-13.54%	-$8,229.81*	$222.21	-13.67%

(b) Income

	Low Income (<= $40,000)			High Income (> $40,000)		
Outcome	Effect	SE	Effect/Base (%)	Effect	SE	Effect/Base (%)
Percentage Points						
College Enrollment	-0.058	0.100	-0.28%	-0.714*	0.119	-1.47%
College Enrollment: Out of State	-0.08	0.114	-0.65%	-0.881*	0.127	-2.79%
College Enrollment: Four-Year	0.008	0.111	0.10%	-0.529*	0.134	-2.19%
College Enrollment: Non-Public	0.019	0.052	0.62%	-0.199*	0.057	-2.41%
College Quality: > $20,000	-0.013	0.095	-0.07%	-0.688*	0.117	-1.49%
College Quality: > $30,000	0	0.085	0.00%	-0.521*	0.109	-1.47%
College Quality: > $40,000	0.017	0.053	0.58%	-0.188*	0.065	-1.62%
Earnings > 0	0.023	0.082	0.03%	0.202*	0.054	0.23%
Earnings > $10,000	-0.103	0.101	-0.36%	0.241*	0.081	0.71%
Dollars						
Earnings	-$41.44	$26.63	-0.56%	$13.79	$19.54	0.15%
College Quality: Alumni Earnings	-$4.34	$21.65	-0.02%	-$148.50*	$28.42	-0.58%
Family Income	-$4,450.52*	$224.12	-14.89%	-$9,790.19*	$217.71	-13.65%

(c) Financial Wealth (All with Income > $40,000)

	Low Wealth (Interest <= $500)			High Wealth (Interest > $500)		
Outcome	Effect	SE	Effect/Base (%)	Effect	SE	Effect/Base (%)
Percentage Points						
College Enrollment	-0.713*	0.121	-1.57%	-0.624*	0.205	-0.93%
College Enrollment: Out of State	-0.88*	0.130	-2.97%	-0.738*	0.222	-1.69%
College Enrollment: Four-Year	-0.494*	0.119	-2.28%	-0.519*	0.238	-1.36%
College Enrollment: Non-Public	-0.163*	0.054	-2.22%	-0.321*	0.143	-2.37%
College Quality: > $20,000	-0.68*	0.119	-1.59%	-0.638*	0.207	-0.99%
College Quality: > $30,000	-0.48*	0.104	-1.50%	-0.638*	0.205	-1.17%
College Quality: > $40,000	-0.139*	0.058	-1.46%	-0.375*	0.176	-1.61%
Earnings > 0	0.13*	0.055	0.15%	0.636*	0.133	0.74%
Earnings > $10,000	0.163*	0.079	0.46%	0.688*	0.164	2.55%
Dollars						
Earnings	-$8.75	$19.84	-0.10%	$139.17*	$34.17	1.80%
College Quality: Alumni Earnings	-$135.04*	$26.56	-0.55%	-$186.02*	$55.90	-0.60%
Family Income	-$9,319.11*	$223.98	-13.61%	-$12,837.68*	$279.13	-14.21%

Notes: (*) indicates significance at 5% level. Presents DD estimates and standard errors using differences across outcomes for children of Layoff and Survivor parents, and across periods 1 and -1. Effect/Base uses base mean outcomes before events occur.

I.F.2 Family Income and Wealth

I now estimate effects separately for higher- and lower-income families. Table 1.5.b displays estimated treatment effects $\hat{\beta}_{1,-1}$ on other outcomes for families with incomes above and below \$40,000. There are no significant effects on low-income families, whereas all the college and labor supply outcomes show significant effects in the expected directions for higher-income families.[23]

Figure 1.7 examines this pattern in more detail. Figure 1.7.a displays treatment effects $\beta_{1,-1}$ of fathers' layoffs on family income, grouped by income level prior to events. Layoffs reduce income levels by more in higher-income families.[24] Figure 1.7.b displays corresponding treatment effects $\beta_{1,-1}$ on child college enrollment, for the same income groups. Treatment effects are close to zero at the lowest incomes, rise steadily, and may begin to decrease at the highest incomes.

[23] There is a concern that this pattern could arise spuriously from weaker father-child links at lower incomes, even within the constraints imposed by my parent-child matching algorithm. To examine this I restricted the sample to children only ever claimed by one adult, that adult being the father, rather than my normal restriction of only being claimed by at most one male and one female adult. This did not change the pattern displayed in the figure. I also checked the probability of being claimed by this one adult at ages before 18 across income levels, and found that this probability is 85% at the very lowest income levels, but quickly rises above 90% for incomes over \$10,000. Therefore even low-income children only claimed by one adult–the father receiving the layoff–most of whom are claimed almost every possible year by that one adult, do not exhibit an enrollment response to father's layoff.

[24] Note that DD treatment effects at base incomes far from the mean exclude mean-reversion due to the use of Survivors.

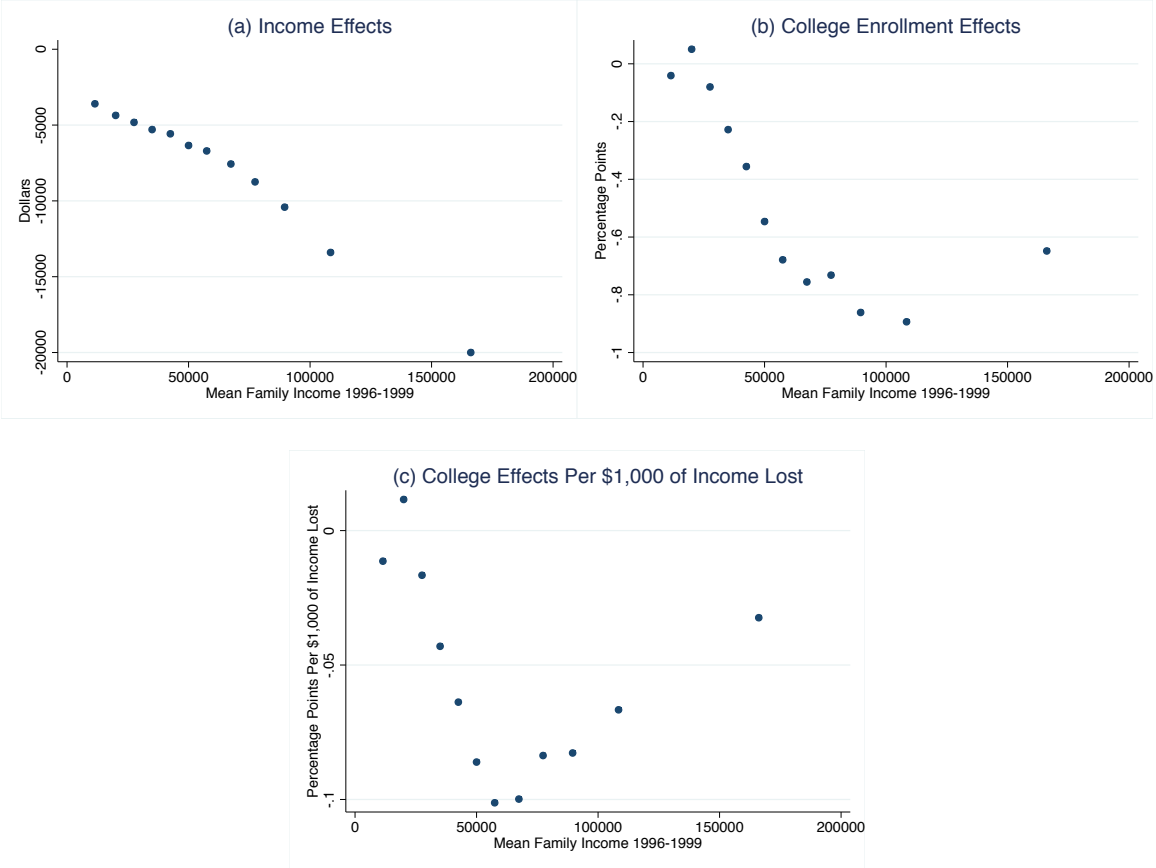


Figure 1.7

Treatment Effects by Family Income

Notes: All panels plot estimated treatment effects by mean pre-tax family income over 1996-1999. Panel (a) plots short-run DD treatment effects $\boldsymbol{\beta_{DD}}$ from equation (3) estimated on total post-tax family income. Panel (b) plots analogous treatment effects on annual college enrollment ages 18-22. Panel (c) plots the treatment effects in Panel (b) divided by the treatment effects in Panel (a), where the income effects in Panel (a) have been divided by $1,000 for ease of interpretation as "Effect on enrollment per $1,000 of family income lost due to layoff."

Figure 1.7.c displays the college effects as percentage points of enrollment per $1,000 of income lost, and reveals a striking U-shaped pattern of treatment effects. This U-pattern is unlikely to arise if layoffs mainly affect children through non-income factors such as father's becoming depressed after layoffs. In contrast, an income-loss channel provides a clear explanation. At low incomes, children do not rely heavily on parents to finance college. In the Sallie Mae data, children with family incomes below $35,000 finance only 19% of college expenses out of parental resources (loans, income, and savings), compared to 41% for families with incomes between $35,000 and $100,000, and 61% for families with incomes above $100,000 (Sallie Mae 2011). Low-income children make up for the difference with greater financial aid, student loans, and student earnings. They also attend lower-cost colleges.

Income losses can also explain the decline in treatment effects as incomes continue to rise past $60,000. There are two leading explanations. The first is that families view college as an investment and face liquidity constraints, as in Becker (1994). The market for private student loans was active in the U.S. over the sample period, suggesting that subsidized federal Stafford loans did not fully meet demand. The interest rates on private loans for college–for the subset of students who qualified–were often much higher than interest rates on collateralized debt such as home mortgages (Delisle 2012). Under liquidity constraints, low-income parents allocate transfers to children in the form of human capital investments. As incomes, transfers, and human capital investments rise, the return on human capital declines to the interest rate on financial assets, at which point parents allocate marginal transfers in forms other than human capital. This means that as incomes rise, the marginal change in transfers from parents to children is less likely to reduce spending on college.

The second reason why effects on college might decline at higher incomes is that families view college partly as a consumption good, and spending on this good becomes a smaller fraction of total spending at higher incomes, e.g., the Engel curve in logs declines with income (Mulligan 1997). For example, a middle-income family that spends 20% of its budget on college will reduce college spending by $20 out of a $100 income loss, while a family that spends 10% of its budget on college wil reduce college spending by only $10 out of a $100 income loss. For college, a natural explanation for declining Engel curves is that each child only needs to enroll in one college, and

tuition is bounded by institutions.[25] Therefore the fact that college becomes less sensitive to income shocks at higher incomes cannot distinguish between investment and consumption mechanisms.

The ideal test to distinguish consumption and investment mechanisms would be to vary current and permanent income losses from layoff. Unfortunately, these two variables are too highly correlated to identify their separate effects. A more feasible test is to examine treatment effects separately for families with high and low financial wealth, as in Zeldes (1989). For this exercise, I restrict the sample to families with pre-layoff incomes above $40,000. Table 1.5.c shows effects of parental layoffs on various child outcomes for families with pre-layoff interest income above and below $500,[26] corresponding to an asset cutoff of about $10,000-$25,000 if interest rates on savings are 2-5%. The effects on child college outcomes in these two groups are not statistically different on an absolute or per-dollar basis.[27] These results provide no evidence of liquidity constraints among middle-to-high-income families, and suggest that parents at these income levels view marginal college expenditures as consumption.[28]

I.F.3 Predicted Earnings Losses

I now explore an additional source of evidence on whether income losses explain the main effects. I first explore one measure of economic vulnerability to layoff: father's earnings share. Earnings losses of fathers reduce family income by more, proportionally, when fathers earn a larger share of family income prior to layoff. This observation suggests that if income losses are driving the effects on children, then effects on children should increase in father's earnings share. Father's earnings shares, however, are not randomly-assigned. The two most important components of family income are father's earnings and mother's earnings. As father's earnings increase, family socioeconomic status (SES) rises, and fraction of income lost from the father's layoff should also rise. As mother's earnings inrease, family SES rises, but now fraction of income lost from the father's layoff should *fall*. By examining these two sources of variation separately, I estimate effects of proportional

[25] There is some evidence that Engel curves decline with family income in the NPSAS data.

[26] This is about the 80th percentile of interest income for families with incomes over $40,000 in my sample.

[27] Different cutoffs for interest income from $0 up to $3,000 do not change the pattern described here, although confidence intervals get wide as the cutoff gets higher.

[28] Anecdotally, many middle-income parents in the U.S. offer to cover children's college costs if children attend lower-amenity, lower-cost state or community colleges, but only a fraction of costs at higher-amenity, higher-cost private colleges. This suggests that parents may view the benefits of more expensive colleges as a form of consumption, and therefore reduce spending on this good when permanent income falls, consistent with the findings.

income losses from layoff that should be biased in opposite directions by confounding variation in family SES.

Define father's pre-event earnings share in period $k = -1$ as $\omega \equiv \frac{W_{-1}^{dad}}{W_{-1}^{dad}+W_{-1}^{mom}}$. I first divide the sample into ω bins.[29] I then reweight these bins by mother's earnings W_{-1}^{mom} in order to isolate variation in ω from father's earnings W_{-1}^{dad}, or vice versa in order to isolate variation in ω from mother's earnings W_{-1}^{mom}. For this exercise I restrict the share groups to a range with enough observations to reweight them on W_{-1}^{dad} or W_{-1}^{mom}, and to relatively high incomes due to the finding above that children's college decisions appear unresponsive to layoff in low-income families.

Figure 1.8 implements this exercise using variation in mother's earnings. Panel (a) shows this variation. As father's earnings share rises from 70% to 100% on the x-axis, mother's earnings fall by \$30,000, while average father's earnings remain constant. Panel (b) plots the effect of layoffs on family income by earnings share. The 30 points of earnings share variation yields an additional 6 percentage points of income loss. Panel (c) plots the effect of layoffs on child college enrollment by earnings share. Enrollment declines more in higher-share groups that experience larger proportional income losses from layoff.

[29] I do not scatter the points in this graph because it is not possible to obtain equally-sized bins, largely because there is a large mass of mothers at zero earnings.

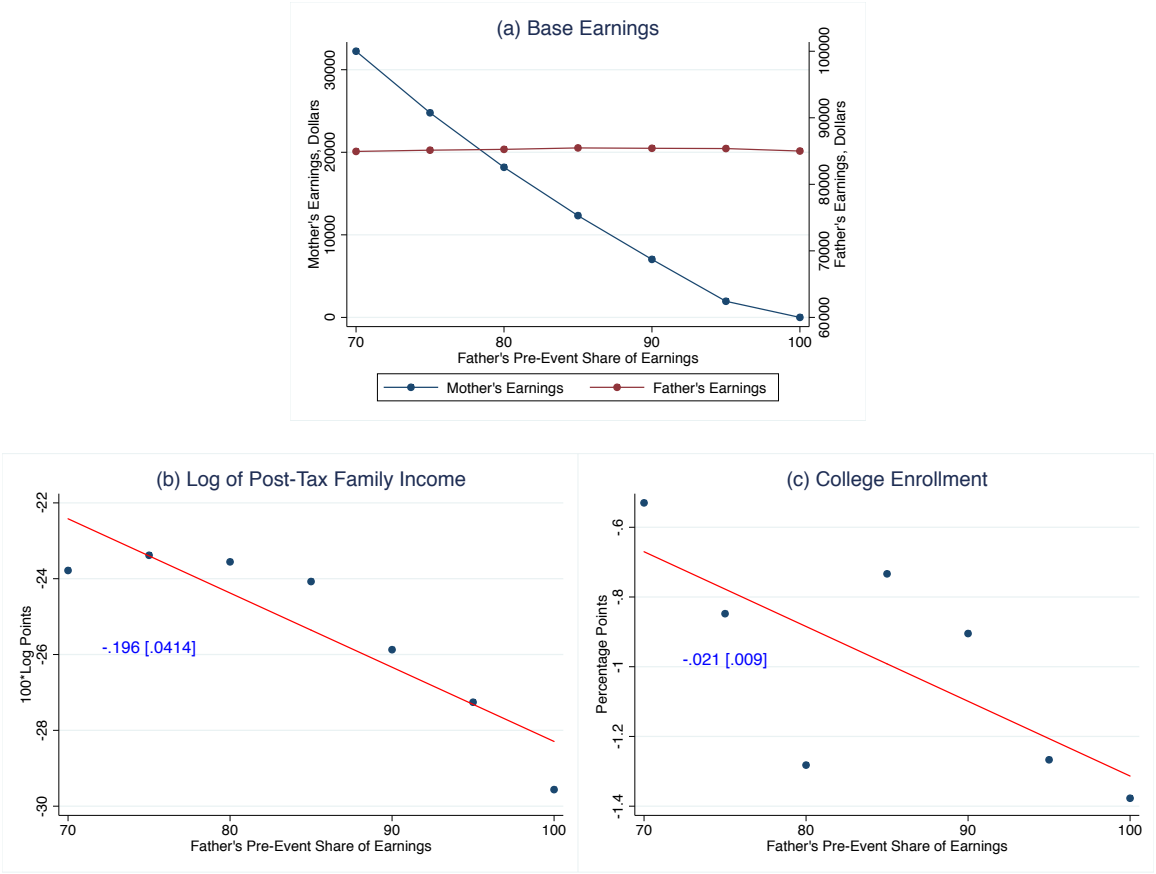


Figure 1.8

Treatment Effects by Father's Earnings Share: Variation from Mother's Earnings

Notes: All of these figures restrict father's earnings in period -1 to lie in [$60,000, $160,000] and plot variables by father's period -1 earnings share, where earnings share groups have been reweighted to match the share=70% share group in father's earnings in -1. Panel (a) plots mother's and father's period -1 earnings levels in this reweighted data. Panels (b) and (c) plot estimated short-run DD treatment effects from Equation (3) on the log of family income and the level of college enrollment, respectively. Regression lines depict slope of estimated coefficient from OLS regression of these treatment effects on father's pre-event earnings share.

Figure 1.9 repeats this exercise using variation in father's earnings. Now as the shares increase from 40% to 60%, father's earnings rise from $30,000 to $65,000, while average mother's earnings remain constant. Whereas family SES declined in father's earnings share in Figure 1.8, it now rises in father's earnings share. Despite the different source of earnings share variation that moves family unobservable characteristics in the opposite direction, once again college enrollment declines more in higher-share groups that experience larger proportional income losses from layoff.

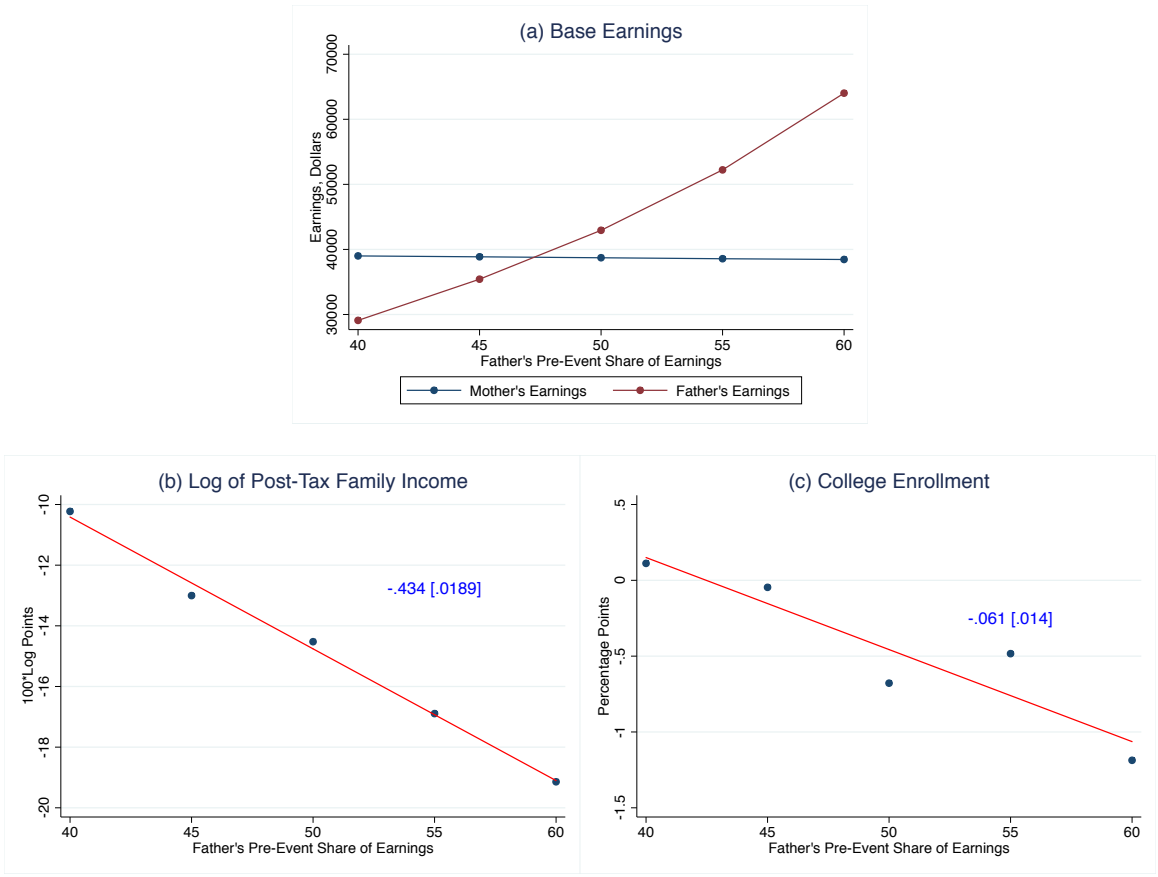


Figure 1.9

Treatment Effects by Father's Earnings Share: Variation from Father's Earnings

Notes: All of these figures restrict mother's earnings in period -1 to lie in [$30,000, $50,000] and plot variables by father's period -1 earnings share, where earnings share groups have been reweighted to match the share=50% share group in mother's earnings in -1. Panel (a) plots mother's and father's period -1 earnings levels in this reweighted data. Panels (b) and (c) plot estimated short-run DD treatment effects from Equation (3) on the log of family income and the level of college enrollment, respectively. Regression lines depict slope of estimated coefficient from OLS regression of these treatment effects on father's pre-event earnings share.

These figures can be interpreted as instrumenting for proportional income losses with father's earnings share. The implied Wald estimator that instruments for proportional income losses with mother's earnings variation in Figure 1.8 is .12, while the analogous Wald estimator for father's earnings variation in Figure 1.9 is .14. The Wald estimator obtained on the full sample of layoffs by instrumenting for proportional income losses with a dummy for layoff is 0.1. These estimates therefore provide no evidence to reject the hypothesis that income losses account for the full treatment effects of layoffs on children's college enrollment.

I have focused on father's earnings shares as one factor that predicts income losses from layoff, in order to show how separate variation in mother's and father's earnings can address concerns about endogeneity of predicted income losses. However, this approach only exploits a small fraction of the information available to predict income losses from layoff, and mother's earnings only provide variation in proportional income losses, not absolute income losses. It is possible to obtain much more precise predictions of father's earnings losses, and variation in absolute income losses, by exploiting all pre-event information about fathers. Define a father's proportional earnings loss around an event as $L_{i,g} = \frac{W_{i,g,k_1} - W_{i,g,k_2}}{W_{i,g,k_2}}$ where $k_1 > 0 > k_2$ as before, and define a vector of pre-event variables X_{i,k_2}. These pre-event variables include information about the father's industry, firm, location, wife, and other demographics.[30] I generate comparable earnings loss predictions for all fathers as follows. I first regress $L_{i,T}$ on X_i restricting to the Layoff sample, then separately regress $L_{i,C}$ on X_i restricting to the Survivor sample. This yields estimated coefficient vectors $\hat{\pi}_T$ and $\hat{\pi}_C$, respectively. I then calculate predicted earnings losses under realized and counterfactual events, yielding $\hat{\pi}_T X_{i,T}$ and $\hat{\pi}_C X_{i,T}$ for Layoff fathers and $\hat{\pi}_T X_{i,C}$ and $\hat{\pi}_C X_{i,C}$ for Survivor fathers. I then group fathers by the difference between these two predictions $\hat{D}_i$, where $\hat{D}_i = (\hat{\pi}_C - \hat{\pi}_T) X_i$ for all fathers. These differences capture fathers' vulnerability to earnings losses from layoff, excluding mean-reversion and other movements in earnings that would happen within X_i groups, even if layoff were not experienced.

Figure 1.10 presents the results from this exercise, where $W_{i,g,k_2} \equiv W_{i,g,-1}$ and $W_{i,g,k_1} \equiv \frac{1}{5}\sum_{j=1}^{5} W_{i,g,j}$, or average earnings one year to five years after events.[31] Figure 1.10.a graphs total post-tax family income losses $\hat{\beta}_{1,-1}$ by this measure of father's predicted earnings loss $\hat{D}_i$. An

[30] Note that $L_{i,g}$ is large across the entire income distribution, and therefore relies on different variation from that explored in the income cuts displayed in Figure 7.

[31] Results are similar for other definitions of post-event earnings W_{i,g,k_1}.

additional percentage point of father's earnings loss increases the loss in family income by $618. Figure 1.10.b graphs the college enrollment decline $\hat{\beta}_{1,-1}$ against the father's predicted earnings loss $\hat{D}_i$. An additional percentage point of father's earnings loss increases the decline in college enrollment by .04 percentage points. Using predicted earnings losses to instrument for income losses yields a Wald estimator for the effect of family income on college enrollment of 0.07 (standard error .012) percentage points of enrollment per $1,000 of income. This is exactly the slope required for income losses to "explain" the entire treatment effect of layoffs on children, and is consistent with the results obtained using father's earnings share variation.

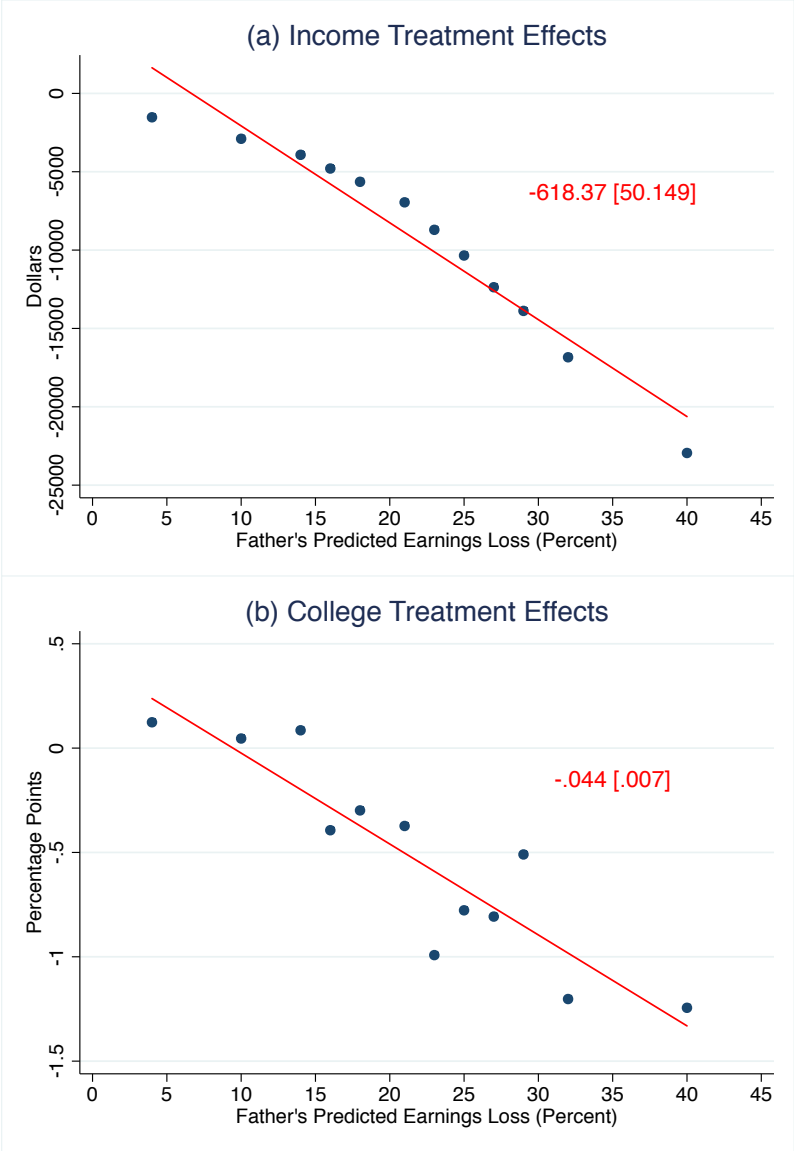


Figure 1.10

Treatment Effects by Predicted Percent Earnings Loss

Notes: Figure 1.plots treatment effects by father's predicted percent earnings loss, here defined as the percent change in income from period -1 earnings to the mean of earnings from period 1 to period 5. Predicted percent earnings loss is constructed from separate regressions of actual earnings losses on pre-event covariates in the Layoff and Survivor samples, where the final prediction is the difference between the predicted values for each observation using the estimates from the Layoff and Survivor regressions. Observations are grouped into 12 predicted earnings loss bins that have approximately equal numbers of Layoff fathers. Panels (a) and (b) plot estimated short-run DD treatment effects using Equation (3) on the level of post-tax family income and the level of college enrollment, respectively. Regression lines depict slope of estimated coefficient from OLS regression of these treatment effects on father's predicted earnings loss.

I.G Discussion of Results

The results above provide strong evidence that parental layoffs occurring just before college decisions have small impacts on college outcomes, despite their large impacts on parental resources. One reason for this might be that children adjust other margins instead of altering investments. While this may be true, adjustments are also small for every alternative margin I observe including college quality, type, distance from home, and child earnings. To understand why adjustments are small at a more basic level it is useful to examine how children actually pay for college. Data on college finance are available in Sallie Mae's (2011) "How America Pays for College" survey and in the Department of Education's National Post-Secondary Student Aid Study.

The average annual cost of college in the U.S. over this study's sample period is about $20,000, including tuition, fees, room and board. Average parent borrowing and transfers finance about $10,000 of these costs. Layoffs reduce permanent family income by about $6,000 and transitory income by about $10,000. A key parameter for predicting effects of this income loss is the fraction of marginal family income that parents would allocate to college spending if their child attended college. This parameter is the conditional-on-college parental Engel curve in college spending. While there are no existing causal estimates of this parameter, a benchmark can be obtained easily from cross-sectional data. Surprisingly, in both Sallie Mae and DOE data, the slope of this line is very flat: about .02 to .05. This suggests that a loss of family income between $6,000 and $10,000 will reduce average parental contributions for college by $120-$500. Deming and Dynarski (2010) review the financial aid literature and conclude that $1,000 of salient, easily-obtained financial aid raises enrollment by about 3 percentage points. This is likely an upper bound on the effect of $1,000 in parental contributions, because children may not view parental contributions as "free money," and parents do not necessarily require children to spend transfers on college. Based on these numbers, a plausible upper bound on the effects of a father's layoff on college enrollment is 0.36-1.5 percentage points. The effect of layoffs I actually find is 0.43 percentage points. In this context, small causal effects of layoff on enrollment appear much less surprising. This reasoning also suggests, again, that income losses are sufficient to explain the full effect of layoffs.

These calculations strengthen the lessons discussed above regarding Figure 1.1.a: financial aid most likely is a much more effective way to increase college enrollment than transfers to parents

with teenage or college-age children. To clarify this I convert the effect of financial aid into a metric that is comparable with the effect of family income: percentage points of enrollment gained per $1,000 of government spending. Average enrollment at age 18 in the U.S. is about 40%. An offer of $1,000 of additional, non-crowded-out financial aid would be predicted to increase this number to 43%. This would cost, on average, $430. The enrollment gain per $1,000 dollars actually spent would therefore be closer to 6 percentage points. In contrast, if layoffs affect parental college spending through transitory income, $1,000 of government spending will raise enrollment by 0.04 percentage points, and if layoffs affect parental college spending through permanent income, $1,000 of additional government spending will raise enrollment by only .007 percentage points. This means that on the margin in the U.S. today, spending on financial aid is 150-850 times more effective at raising enrollment than unanticipated cash transfers to parents with college-age children. It is possible that *anticipated* cash transfers in late childhood, which parents may smooth into child inputs at earlier ages, have large impacts on college enrollment and other important child outcomes. But both this smoothing behavior and the effects of the earlier child inputs must be larger than any existing evidence suggests in order to reach a different conclusion.

A simple, reduced-form policy calculation serves to make this point more vivid. The Dependent Exemption and the non-refundable portion of the Child Tax Credit transferred $60 billion to middle- and high-income parents in 2008. Consider a policy that re-allocates this revenue to financial aid for children aged 18-22 with below-median parental income. I estimate that these income losses after age 12 have virtually no impact on enrollment, and estimates in Dahl and Lochner (2012) and Chetty, Friedman and Rockoff (2011) suggest income effects before age 12 for middle- and high-income families will reduce enrollment by at most one percentage point.[32] The revenue saved would finance $12,000 in annual offered scholarships for every child between ages 18-22 below median parental income, raise enrollment during these ages by 36 percentage points, and close 90% of the entire 2-year gap in total college attainment between children in families with above- and

[32] Dahl and Lochner (2012) estimate that $1,000 of permanent income raises child test scores by 6% of a standard deviation, where most of this is driven by children in lower-SES families and therefore represents an upper bound on the effect of foregone non-refundable tax credits that mostly benefit middle- and upper-income families. Chetty, Friedman and Rockoff (2012) find that a one standard deviation gain in student test scores generates a 5 percentage point gain in college enrollment and a 10% gain in future earnings. Suppose each family losing income has on average 2 children. Then families on average lose $2,000 of income from the policy change. Under the strong assumption that test score changes produced from teacher quality and family income map into long-term outcomes identically, the income loss produces a 0.6 percentage point reduction in future enrollment, and a 1.2% decline in future earnings.

below-median incomes.[33] This exercise is based on very strong assumptions, but it illustrates that large gains from revenue-neutral shifts in the composition of government spending on children are plausible.

If parental college spending, like financial aid, does have large per-dollar impacts on child college outcomes, why don't parents devote more income to helping children pay for college–why is the Engel curve so flat? Even at high incomes, parents leave children to finance much of college with student loans, earnings from work and low consumption, when apparently a few thousand extra dollars of help for a few years could compel many of these children to attend college. It is possible that parents believe children who are unwilling to endure these costs have little to gain from college. But with an average lifetime earnings premium for college graduates approaching 60% (Goldin and Katz 2008) over the sample period, it seems worthwhile for altruistic parents to bribe children into obtaining a degree (as in Weinberg 2001). Another explanation is that many child inputs at earlier ages, such as a mother's health while pregnant, high-quality neighborhoods and comfortable home environments, are within-family public goods that benefit parents and children simultaneously. In contrast, college is more of a private good that primarily benefits children. College spending may therefore place larger demands on parental altruism. This theory predicts that parental income at earlier ages will have larger impacts on long-term child outcomes even if the per-dollar productivity of early childhood inputs are identical to the per-dollar productivity of college investments.

Do effects of parental layoffs on college outcomes, in themselves, justify substantial tax and social insurance advantages for parents over non-parents? Under strong assumptions, I estimate the NPV of the earnings loss imposed on children by parental layoffs to be \$1,000-2,000.[34] This is small compared to the transfers that would be required to replace lost family income from layoff while children are in college.

[33] There are approximately 3 million children per cohort, and therefore 15 million children between ages 18-22 each year. \$60 billion permits \$8,000 of annual realized spending per child below median parental income. Define affordable offered aid G, affordable realized aid E, the causal effect ϕ of \$1,000 of offered financial aid on enrollment, and base enrollment R among the aid-eligible population. Under the strong assumption that ϕ holds true at levels of G substantially larger than most variation used to estimate ϕ in existing studies, and that ϕ is true for all ages 18-22, these variables are related by the identity $E = (R + \phi G)\, G$. For $R = 0.3$, $\phi = 0.03$ (Deming and Dynarski 2010, and consistent with results in this paper), and $E = \$8,000$, this yields $G \approx \$12,000$, an annual enrollment gain of $\phi G \approx .36$, and a gain in total college attainment of $5\phi G \approx 1.8$ years. Future research on the causal effects of such a large, means-tested financial aid program would be very useful.

[34] I find that layoffs reduce total college attainment during ages 18-22 by 0.016 years. If each year of college raises earnings by 10% and lifetime earnings are \$500,000-\$1,000,000, then parental layoffs reduce children's future earnings by \$800-\$1,600. An alternative way to reach a similar conclusion is to impute the future earnings loss directly from the 0.35% decline in alumni-earnings-based college quality.

I.H Conclusion

This paper has demonstrated that large, unanticipated family income shocks during late childhood have small, adverse effects on children's college enrollment and choice of college. The research design uses fathers who are laid off in the near future as a control group for fathers who are laid off in the near past, and a reweighted sample of fathers who experience "control events" to remove time trends among laid-off fathers. The design identifies selection-on-unobservables into layoff (γ) separately from treatment effects (β_{DD}) and reveals that selection of parents into layoff both within and across firms is important, even when conditioning on a rich set of family background characteristics. This finding bears on a broad range of empirical applications in which it is difficult for the researcher to observe a dependent variable for the same individual both before and after a firm-level treatment takes place.

I precisely estimate that an unanticipated $1,000 decrease in permanent income due to a father's layoff reduces children's enrollment by 0.18%. I find similarly small, adverse effects on several measures of college quality and distance of college from home. I also find that all of these effects are *smaller* for low-income children. Using supplementary data, I show that the smaller effects I obtain after controlling for selection-on-unobservables into layoff are more consistent with how children finance college in the U.S. Layoffs most likely reduce parent college spending by only $100-500. The fact that layoffs have any effect at all on children's enrollment therefore *reinforces* findings in the financial aid literature that children respond strongly to perceived college costs.

I provide a variety of arguments consistent with the interpretation of layoffs as family income shocks. An income loss channel offers a simple explanation for the U pattern of treatment effects on enrollment by family income. An income loss channel also accords with treatment effects being larger in families that rely more heavily on father's earnings, and larger in families with fathers predicted to lose more earnings from layoff. Finally, I show that reductions in college enrollment and quality following layoff most likely reflect lower family consumption rather than liquidity constraints on investment.

This paper contributes to the debate over how to improve long-term outcomes of disadvantaged children. Programs that explicitly target children and parents account for about $300 billion or 10% of annual federal spending in the U.S. (Isaacs *et al* 2011). Income transfers to parents account for

half of this spending; subsidies for specific child inputs such as education and health care account for the other half. The findings here show that policies designed to alleviate credit constraints for parents of teenage and college-age children are unlikely to have a significant impact on college outcomes, especially for low-income children. The findings are also consistent with existing evidence that salient, easily-obtained financial aid has much larger impacts (Deming and Dynarski 2010, Bettinger *et al* 2009). These considerations suggest that revenue-neutral shifts in the composition of government spending on children—out of some parental income transfers and into specific child input subsidies—may have much larger impacts on future generations than previously thought.

Appendix 1: Matching Algorithm and Effects on Sample

I first discuss the general logic of the match of fathers and mothers to children and then document the exact routine employed. Linking parents and children in IRS data for my event study and event-age study designs requires care for several reasons. Marital status and children are only reported by filers, and filing is reduced by layoff. Therefore it is important to use information prior to layoffs to match parents and children, a rule I follow with one exception, discussed below. All matches of fathers with children rely on claims from 1996-1998, giving a buffer of two years before the first layoff can occur, in 2000. I also restrict to claims in years before a child turns 18, because after that age claims depend endogenously on child college outcomes for eligibility reasons. Over 90% of matches occur in the first available year, 1996, while virtually all the rest are made in 1997[35]. About 10% of children are only claimed by mothers and therefore excluded from my sample. An additional 25% of children are either never claimed, or claimed by too many different people for my matching algorithm to assign them a single father in all years 1996-2009 with confidence, and therefore removed from the sample. Multiple claimers are a much bigger problem than no claimers, because most low-income parents file taxes and claim children in order to collect large EITC benefits (Athreya *et al* 2010).

Removing children claimed by multiple fathers before age 18, even when a second father claims the child after a first father is laid-off, violates the rule that only pre-layoff information be used in matching children to parents. My match algorithm errs on the side of strong linkages to assure that children are linked with their primary, contempary source of parental support. Measurement error in family linkages can be a minor problem when using parents as background controls in the context of some external treatment, because even a non-contemporaneous parent likely contains a lot of information about fixed characteristics of a child's family background. However, erroneous linkages are a major problem when measuring the effects of *changes* in parental circumstances over time on child outcomes. Even if all of a child's claiming fathers have highly-correlated fixed characteristics, changes in their time-varying characteristics are likely far less correlated. Therefore I err on the side of excluding children claimed by more than one father to assure I have strong parent-child linkages.

[35] The claims data are mostly missing in 1998-2000. It is therefore reassuring that many more kids are claimed for the first time in 1996 than 1997, because this suggests that the missing data only cause a tiny fraction of missed linkages.

This restriction eliminates nearly 20% of children ever claimed in IRS records. Unsurprisingly, children claimed by multiple fathers or no father have much lower college enrollment than children claimed by one father (note that college enrollment is observed for all children, both matched and unmatched). To the extent we think income matters more for children in lower-SES families, my estimates may be smaller than estimates for the full population of children. My estimates may also be smaller if layoffs affect children through mechanisms that correlate with divorce and remarriage (the most likely path to multiple claimers) before a child turns 18.

Figure 1.A1.1 shows two simple validations of my algorithm for matching parents and children. It plots the pre-tax income distribution for a random sample of children in my IRS data who are age 14-16 in 2001, compared with two samples of age 14-16 children drawn from the 2001 American Community Survey (ACS). My sample selection criteria cannot be validated exactly in ACS data because it relies on the time dimension of my panel data. I therefore use two ACS samples with income distributions that I expect to "bracket" that of the IRS sample. The main issue is that children in my data are in households that were headed by men at some point during 1996-1998, several years before the year of observation, 2001. These households are higher-SES than average Census households due to their male-headed status, but lower-SES than Census households headed by men in 2001. Figure 1.A1.1.a confirms this intuition when these three income distributions are normalized into PDF's. The income distribution resulting from my linked sample looks very close to what would be expected from ACS data. Figure 1.A1.1.b makes another point: the Average sample of children is smaller (using appropriate sampling weights) than the Census male-headed sample. This is because I exclude children who are claimed by more than one father before age 18, and children who are never claimed on tax returns.

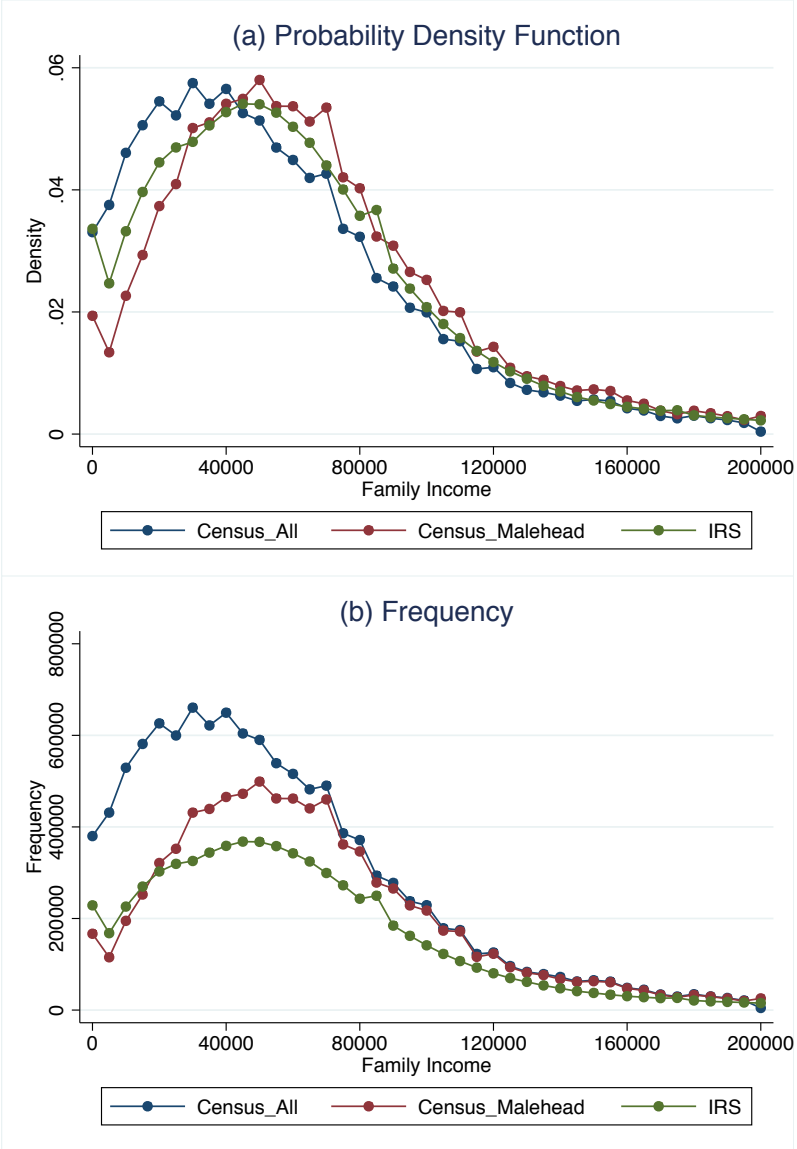


Figure 1.A1.1

IRS Linked Sample vs Census

Family Income Distributions of Children Age 14-16 in 2001

Notes: IRS linked sample is a random sample of the children age 14-16 in 2001 who I match to fathers in 1996-1998. Census_All is a random sample of children from the ACS age 14-16 in 2001. Census_Malehead is a random sample of children from the ACS age 14-16 and residing in male-headed households in 2001. Income is year 2001 pre-tax family nominal income. Panel (a) presents the probability density functions for the income distributions of these three samples. Panel (b) presents the corresponding histograms for these three samples.

The algorithm is as follows.

1. Make a list all unique pairs of children and claimers in every year in the sample, 1996-2009. Each individual is indexed by a unique identifier. About 95% of individuals currently in the US who were children during this sample period are linked to at least one claimer. Some of the remaining 5% may have arrived as immigrants after age 18.

2. Restrict this list to rows in which the child is under 18, because claims beginning in 18 are typically only valid conditional on college enrollment.

3. Get the sex of all claimers and the unique identifier of each claimer's spouse, if any, in every year. In each year call the claimers "PE's" for "primary earnings," and their spouses "SE's" for "secondary earners."

4. Case 1: Child has only one PE claimer (63.7% of children)

   -Restrict to SE's who claim child largest number of times.

   -If multiple SE's, break tie by selecting SE who claims child first

5. Case 2: Child has exactly one male PE claimer and one female PE claimer (11.4% of children)

   -Assign child this man and this woman as mother and father

6. Discard remaining children (20% of children)

Appendix 2: Firm Closure Sample: Details

I here briefly discuss the sample restrictions of the closure sample I construct in IRS data, which are based on those in OPS.

Firms can only enter the sample if they employ at least 30 workers at time of closure.

One restriction requires that fewer than 35% of workers experiencing closure at T at a particular firm be working at the same firm in a future year, and is intended to remove re-organizations mistakenly identified as closures. This restriction eliminates 45% of candidate firm closures.

Other restrictions require at least two years of zero UI and two years of tenure at the closing firm. The tenure restriction eliminates 35% of workers and the zero-UI restriction eliminates 15%

of the remaining workers. To identify closure at T under these restrictions, we need to confirm zero employment at that firm in T+1, no excess bunching of displaced workers at the same firms in T+2, and the tenure and no-UI restrictions in T-1 and T-2. It is also important to note that many spurious closures arise if less than two years are allowed for late updates of the W2 earnings data in IRS records. Imposing all these restrictions require me to limit my sample to closures to 2001-2007. I also impose a restriction that fathers earn less than $150,000 (2009 dollars) in both of the two years prior to layoff, because there is not enough overlap in this region to adequately reweight Non-Closures to Closures.

The resulting closures sample is a 100% sample of workers displaced by closure who take up UI, combined with a 30% random sample of workers displaced by closure who do not take up UI, with appropriate sampling weights.

Table 1.A2.1 shows number of firms that close and their average size, by year of closure, in my sample. While some fairly large firms do close every year, the vast majority of closing firms are small. This leads to the small average size of closing firms.

Table 1.A2.1: Firm Closures by Year and Size

Year of Closure	Number of Firms	Mean Firm Size
2000	12,894	156
2001	13,164	153
2002	10,679	164
2003	9,652	121
2004	8,966	118
2005	8,979	140
2006	9,688	121
2007	10,910	99
Total	84,932	136

Table 1.A2.2 displays summary statistics for the Closure and Non-Closure samples, and is analogous to Table 1.1. The Closure and Non-Closure samples display similar overall patterns, though smaller declines in child earnings because many fewer cohorts are included in this sample due to the computational demands of identifying closures according to the above restrictions.

Table 1.A2.2: Summary Statistics 1999-2009 for Children at Age 18

Sample:	Closure			Non-Closure		
	Pre-Shock	Post-Shock	% Diff	Pre-Shock	Post-Shock	% Diff
Parent Outcomes						
Father's Earnings	49,373	40,421	-18.1%	50,947	52,080	2.2%
Father Married	0.792	0.798	0.8%	0.811	0.809	-0.2%
Mother's Earnings	21,885	23,041	5.3%	23,370	23,805	1.9%
Child Outcomes						
Enrollment	0.346	0.368	6.5%	0.393	0.409	3.9%
Earnings	4,128	3,917	-5.1%	3,973	3,862	-2.8%
Freq	35,207	54,031		1,796,769	1,839,753	

Notes: Non-Closure sample is propensity-score reweighted to match Closure sample on observables. "Pre-Shock" includes cells in years before events occur, "Post-Shock" includes cells in years after events occur. "% Diff" is the percent difference between Post-Shock and Pre-Shock columns. Averages pool all available cohorts.

Appendix 3: Identification Using Only Layoffs

In this appendix I develop an estimator that relies entirely on the Layoff sample.

The challenge is to estimate period effects when event-year and cohort both have effects on outcomes that are large relative to period effects, given that these three variables are linearly dependent[36]. This linear dependence makes it infeasible to estimate period effects while controlling for cohort and event-year fixed effects. I first discuss the approach I take intuitively, and then formalize it using the notation developed above.

Estimating treatment effects requires estimation of potential outcomes under non-layoff for the Layoff children after layoff takes place. Above, I use Survivors for this. Here, I use Layoff children prior to realization of layoffs. This is another form of DD estimator. The first difference is the same: the difference between two moments in the Layoff sample on either side of period $k = 0$. Above, the second difference is between the two corresponding moments in the Survivor sample. Here, the second difference is between two moments in the layoff sample, both of which involve $k < 0$. There are typically a number of candidate pre-layoff differences that can be used to estimate the desired potential outcomes. The approach I develop here pinpoints a particular weighted combination of these differences that addresses the problem of confounding event-year and cohort shocks.

Figure 1.A3.1.a displays average enrollment for the Layoff group at age 19 for three cohorts, each plotted by event-age $a_E \equiv a - (t_O - t_E)$. One option would be to pool all of these cohorts into event-age means, but this throws away a lot of useful information. The key information to exploit is that event-age is collinear with event-year for a fixed cohort. Consider the difference $A - B$ for the 1984 cohort. This is a particular "treatment difference," which can be defined as the outcome at an event-year that takes place after the age of the outcome, minus the outcome at an event-year that takes place before the age of the outcome. For cohort 1984, the points $A - B$ reflect outcomes for children with event-year 2002 minus outcomes for children with event-year 2004. The difference $A - B$ therefore contains both the desired difference in outcomes across children with different event-ages, and a confounding difference across children with event-years 2002 and 2004. We would like to estimate the confounding difference across children with these event-years.

[36] The problem occurs when linearly dependent covariates enter a conditional expectation function, and the researcher is primarily interested in effects of a subset of these linearly dependent variabless. Leading examples of this are age, year, and cohort effects in labor economics (Hall, Mairesse and Turner 2005) and age, year, and vintage effects in studies of capital goods (Hall 1971, Berndt, Griliches and Rappaport 1995).

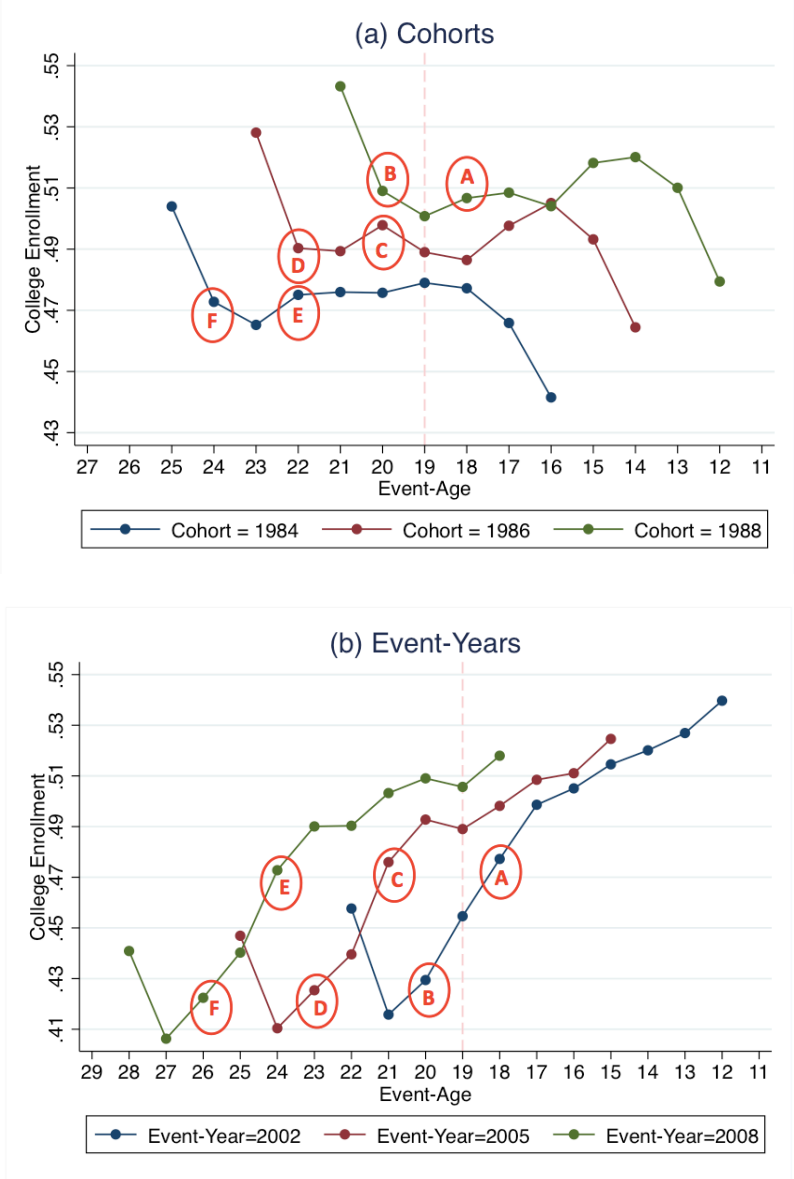


Figure 1.A3.1

Outcomes at Age 19 by Event-Age, Layoff Sample

Notes: Figures plot mean college enrollment for children grouped into cohort by event-age bins in the Layoff sample. Panel (a) displays these group means for three cohorts. Panel (b) displays these group means for three event-years.

There are many ways to estimate this confounding difference. Any cohort for which both of these event-years occur too late to affect the outcome at age 19 provides an estimate of this event-year effect difference. Define a "control difference" as a difference across outcomes for event-years that occur too late to affect these outcomes. Figure 1.A3.1.a presents two such control differences. The control difference $C - D$ uses the 1986 cohort to estimate the difference across event-years 2002 and 2004. The control difference $E - F$ uses the 1988 cohort to estimate the difference across event-years 2002 and 2004. There are many such control differences. Each control difference yields a different double-difference estimator of the treatment effect. One estimator is $A - B - (C - D)$. Another estimator is $A - B - (E - F)$. The unweighted mean of these double-differences provides one estimate of the difference in event-age effects across event-ages 18 and 20, which is the short-run treatment effect of interest.

This estimate relies on the treatment difference in cohort 1984. There are many other cohorts, each of which offers one treatment difference across event-ages 18 and 20. Each cohort uses a different pair of event-years in its treatment difference, and therefore requires a different set of control differences to remove the confounding event-year variation. Each cohort then yields a different "treatment effect," defined as a treatment difference minus the mean of all available control differences that share the event-years used in the treatment difference. I then take the mean of all these treatment effects.

A similar argument holds for treatment differences that occur within event-years, rather than within cohorts. The analogous graph is presented in Figure 1.A3.1.b. I omit the discussion of these estimators to save space; it is conceptually analogous to that just presented. The surprising fact, however, is that the treatment effects that emerge from these two approaches contain independent information. The amount of independent information decreases in the smaller dimension of the outcome-year by event-year matrix. I therefore calculate the complete set of treatment effects and pool them into a single estimate. This approach does count some information multiple times, and therefore overestimates the precision of the final estimate. I ignore this problem.

Write birth cohort as $t_B \equiv t_O - a$. I now rewrite the model in terms of cohort instead of outcome-year. Write the potential child outcome function in terms of a treatment effect, and age interacted with cohort effects and event-year effects, dropping the g subscript because now all observations set $g = T$:

$$y_{a,t_B,t_E} = \alpha + \sum_{j=k_{\min}}^{k_{\max}} \beta_j \cdot I\{k=j\} + \sum_{j=t_B^{\min}}^{t_B^{\max}} \theta_{a,j} \cdot I\{t_B=j\} + \sum_{j=t_E^{\min}}^{t_E^{\max}} \psi_{a,j} \cdot I\{t_E=j\} + u_{a,t_B,t_E},$$

where β_j is the effect of layoff on the outcome, $\theta_{a,j}$ is the effect of cohort j on the outcome at age a, $\psi_{a,j}$ is the effect of selection into layoff in year j on outcome at age a, and u_{a,t_B,t_E} is an error term. A key restriction here is that cohort effects are constant across event-years, and event-year effects are constant across cohorts, within age groups. Without Survivors we have no way to distinguish these interaction terms from event-age effects. This is why the estimates using only Layoffs are much noisier: event-ages are capturing both treatment effects and random cohort by event-year by age interaction shocks.

It is not possible to identify all of the parameters in this model without further assumptions, due to the collinearity $k = t_O - t_E = a + t_B - t_E$. I therefore make an additional selection that period effects not driven by treatment effects are linear:

$$A1 : \beta_j = \phi_{a,0} + \phi_{a,1} k + \sum_{j=0}^{k_{\max}} \lambda_j \cdot I\{k=j\},$$

where $\phi_{a,0}$ and $\phi_{a,1}$ capture the linear trend in period, and λ_k captures treatment effects, assumed to equal zero for outcomes prior to events. This is not a strong additional assumption; it is weaker than the parallel trends assumption $\phi_{a,1} = 0$ used for the main results.

Under these assumptions we identify many different treatment effects using DDs, as previously described. I here characterize the set of these treatment effects. All such DD's consist of one treatment difference that crosses the cutoff where $k = 0$ (e.g.,$a_W = a$, depicted as difference $A - B$), and one control difference that is contained entirely in the untreated region where $k < 0$ (e.g., $a_W > a$, depicted as differences $C - D$ and $E - F$).

Writing event-age in terms of event-year and cohort and fixing age a for simplicity, this set of DD's identifying treatment effects $k = a + t_B^{i_2} - t_E^{i_2}$ years after layoff can be characterized as:

$$\begin{aligned}\Gamma_a\left(t_E^{i_1}, t_E^{i_2}, t_E^{i_3}, t_E^{i_4}, t_B^{j_1}, t_B^{j_2}, t_B^{j_3}, t_B^{j_4}\right) \equiv & E\left[y_a\left(t_E^{i_2}, t_B^{i_2}\right) - y_a\left(t_E^{i_1}, t_B^{i_1}\right)\right] \\ & - E\left[y_a\left(t_E^{i_4}, t_B^{i_4}\right) - y_a\left(t_E^{i_3}, t_B^{i_3}\right)\right] \text{ (DD's in event-age)}\end{aligned} \tag{5}$$

such that

$$1.\ t_E^{i_1} - t_B^{i_1} + 1 \quad < \quad a+1 \leq t_E^{i_2} - t_B^{i_2} \leq t_E^{i_3} - t_B^{i_3} < t_E^{i_4} - t_B^{i_4} \tag{6}$$

(one treatment difference minus one control difference)

and *either* of the following hold:

$$2A.\ t_B^{i_1} \quad = \quad t_B^{i_2},\ t_B^{i_3} = t_B^{i_4},\ t_E^{i_1} = t_E^{i_3}, \text{ and } t_E^{i_2} = t_E^{i_4} \tag{7}$$

(treatment differences removes cohort, control difference removes event-year)

$$2B.\ t_E^{i_1} \quad = \quad t_E^{i_2},\ t_E^{i_3} = t_E^{i_4},\ t_B^{i_1} = t_B^{i_3}, \text{ and } t_B^{i_2} = t_B^{i_4} \tag{8}$$

(treatment differences removes event-year, control difference removes cohort).

To avoid clutter we can re-write $\Gamma_a\left(t_E^{i_1}, t_E^{i_2}, t_E^{i_3}, t_E^{i_4}, t_B^{j_1}, t_B^{j_2}, t_B^{j_3}, t_B^{j_4}\right)$ as $\Gamma_a\left(a_E^{i_1}, a_E^{i_2}, a_E^{i_3}, a_E^{i_4}\right)$, where the assumptions embodied in $\Gamma_a(.)$ are implicit. As stated above for the example, under assumption $A1$ selection terms cancel out in DD's of this nature, and we have:

$$\Gamma_a\left(a_E^{i_1}, a_E^{i_2}, a_E^{i_3}, a_E^{i_4}\right) = \lambda_k. \tag{9}$$

Table 1.A3.1 calculates treatment effects as the unweighted average all these DD's, and is analogous to Table 1.3. Column 1 shows the mean treatment effect for $\pi_a(a-1, a+1)$ by age a for ages $18-22$, as well as a total effect that combines all ages in this range. The estimated effects are very similar to those estimated with Survivors as a control group, but much noisier due to the lack of any way to eliminate event-year by cohort interaction terms.

Table 1.A3.1 Layoffs Only: Causal Effects of Paternal Layoff

Age	College(a-1,a+1) (1)	Income(a-1,a+1) (2)	College(a-1,a+1) Cross-Sectional Prediction (3)	Fraction Causal (4)
18	0.0016	8,420	0.0375	4.4%
	(0.0041)	(819)	(0.0036)	(11)
19	0.0069	8,650	0.0525	13.1%
	(0.0047)	(866)	(0.0053)	(9)
20	0.0061	8,771	0.0498	12.2%
	(0.004)	(737)	(0.0042)	(8)
21	0.0034	8,819	0.0455	7.5%
	(0.0043)	(759)	(0.0039)	(9.5)
22	0.0044	8,868	0.0403	11.0%
	(0.0043)	(858)	(0.0039)	(10.7)
18-22	0.0045	8,664	0.0449	10.1%
	(0.0047)	(816)	(0.0042)	(10.6)

Notes: Column (1) displays the DD estimate of the increase in college enrollment at age *a* from experiencing paternal layoff one year after age *a* compared to one year before age a. Column (2) presents the same estimate for the child's post-tax family income at age *a*. Column (3) multiplies Column (2) by the age-*a* cross-sectional correlation between mean family income 1996-1999 and college enrollment at age *a*. Column (4) displays Column (1) as a percentage of Column (3).

Appendix 4: Linear Differential Trends and Other Robustness Check Results

In this appendix I derive formulas for point estimates and standard errors on a treatment effect estimator that allows for linear differential selection in outcomes with respect to period k for $k < 0$. This is a weaker version of the parallel-trends assumption.

The key parameters are the β_k^T and β_k^C terms; their difference captures the difference in child outcomes around period of father's layoff. These terms are estimated using OLS on Equation (2). I here employ a small amount of new notation for convenience. Write a conditional expectation function for a scalar child outcome Y as

$$E[Y|X] = \beta X,$$

where Y is the child's outcome, β is a K by 1 vector of parameters, and X contains the covariates, including the period terms interacted with type of event (layoff or survival) and controls for event-year and cohort. Let $V_\beta = Var\left(\hat{\beta}\right)$.

First define a 7 by K matrix L_T such that $L_T\beta = \left(\beta_{-7}^T, ..., \beta_{-1}^T\right)$, and similarly define L_C such that $L_C\beta = \left(\beta_{-7}^C, ..., \beta_{-1}^C\right)$, where I have here imposed a cutoff of seven years prior to layoff. Now let $L = L_1 - L_0$, such that $L\beta = \left(\beta_{-7}^T - \beta_{-7}^C, ..., \beta_{-1}^T - \beta_{-1}^C\right)$. This vector contains the points in the pre-treatment region of the graph. We want to estimate a line through these points, i.e., we want to regress these points on a constant and on a linear period trend, where period goes from -7 to -1. Therefore define a covariate matrix

$$Z = \begin{pmatrix} 1 & -7 \\ 1 & -6 \\ ... & ... \\ 1 & -1 \end{pmatrix},$$

and define the parameter vector γ as the least-squares approximation

$$\gamma \equiv \arg\min_a (L\beta - Za)'(L\beta - Za) = \left(Z'Z\right)^{-1} Z'L\beta.$$

We can write the estimator of γ as $(Z'Z)^{-1} Z'L\hat{\beta} \equiv \Omega\hat{\beta}$, and the covariance matrix for $\hat{\gamma}$ as

$\Omega V_\beta \Omega'$.

The target parameter is the estimated difference between the imputed counterfactual outcome under survival and the realized outcome under layoff in period $k > 0$. Define this scalar parameter as $\theta \equiv (\gamma_0 + k \cdot \gamma_1) - \left(\beta_k^T - \beta_k^C\right)$, where we here focus on the case of $k = 1$, the year after layoff. This can be rewritten by defining two matrices $H_0 = (1, k)$ and H_1 such that $\theta = H_0\gamma - H_1\beta$, or

$$\theta = (H_0\Omega - H_1)\,\beta.$$

This neatly writes the target parameter as a linear combination of the original regression of outcomes on period dummies for each group and other controls. We can therefore write the variance of $\hat{\theta}$ as

$$V_\theta = (H_0\Omega - H_1)\,V_\beta\,(H_0\Omega - H_1)',$$

yielding a standard error on $\hat{\theta}$ as $V_\theta^{1/2}$.

Appendix 5: Institutional Non-Filing of the 1098T Form

My results rely on data contained in 1098T forms filed by all Title IV post-secondary institutions. Title IV institutions contain most four-year, two-year, and professional schools in the U.S.. However, recent work by Cellini and Goldin (2012) suggests that 27% of college students are not enrolled in Title IV institutions, and will therefore not receive 1098Ts. In addition, schools are only required to file 1098T forms for individual students who pay any positive dollar amount for tuition, room, board, or other fees, net of financial aid received from the school or other sources.

My analysis data set defines college enrollment as a non-missing 1098T form, and non-enrollment as a missing 1098T form. I have interpreted reductions in 1098T filing for children with recent paternal layoffs as reductions in college enrollment. However, a decline in non-missing 1098T forms could be generated by increased enrollment in non-Title-IV institutions, or by an increase in financial aid for students that pushes their net payments to zero at a school that does not report 1098Ts when not legally required to do so. Therefore, it is possible in theory that my key enrollment decline results represent a switch from Title IV to non-Title-IV institutions, or an increase in 1098T non-filing. The switch into non-Title-IV institutions seems unlikely because

Cellini and Goldin (2012) estimate that these schools are approximately equal in price to Title IV schools, after accounting for subsidies (mainly Pell Grants) at Title IV schools. Therefore the main worry is that layoffs increase 1098T non-filing rather than decrease real college enrollment. While the evidence I present in this appendix suggests 1098T-nonfiling is not driving my results, such a problem would anyway work in favor of my conclusion that layoffs and their associated income declines have at most very small effects on child college outcomes.

To address this concern I first construct an alternative measure of college enrollment in IRS data. I then provide two tests suggesting that the problem is unlikely to drive my main results.

The alternative measure of college enrollment is based on the claiming of children age 19-24. Parents are allowed to claim children ages 19-24 if and only if the child is "permanently and totally disabled" or enrolled full-time at a school[37]. The key features of this rule, for our purposes, are that children can qualify as students if the family pays zero net tuition, and if the school is not a Title IV school. Therefore, conditional on parents filing a tax return, the fraction of parents that claim a child age 19-24 represents a potential alternative measure of college enrollment that can validate findings with 1098T-based enrollment.

Figure 1.A5.1 plots the two measures of college enrollment for children at age 19 in 2002 by mean three-year family income. I restrict to children in 2002 to facilitate comaparison with statistics in the NLSY 97 for cohorts 1979-1982 in Bailey and Dynarski (2011, Figure 1.2). I restrict to age 19 because after age 19 children gradually start to claim themselves as dependents. While parent claims continue to track 1098T-based enrollment for higher-income families, the mechanical decrease in levels makes them less useful as a gauge of total enrollment.

[37] See instructions for the 1040 online at <http://www.irs.gov/instructions/i1040a/ar01.html>.

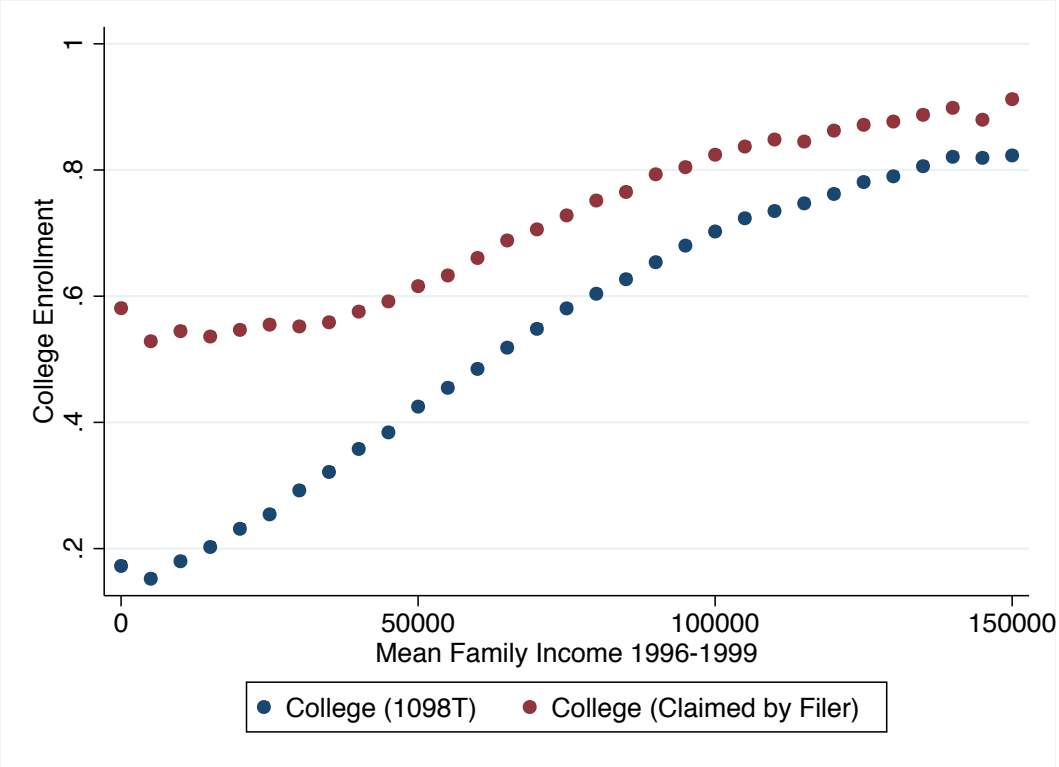


Figure 1.A5.1

College Enrollment at Age 19 in 2002 by Family Income

Two Alternative Measures

Notes: Figure 1.plots two measures of college enrollment in the Survivor sample. “College (1098T)” uses an indicator variable for receipt of a 1098T form by a student. “College (Claimed by Filer)” uses an indicator for whether a child is claimed by an adult (not necessarily the linked father).

The first pattern is that, for richer children, college enrollment based on claims roughly tracks college enrollment based on 1098Ts. About 20% more high-income children are enrolled in college using the claims measure. This is slightly less than the fraction of students estimated to be enrolled in non-Title-IV colleges in Cellini and Goldin (2012). The smaller figure here could be due to the small fraction of 19-year-olds who claim themselves and hence cannot be claimed by their parents.

The second pattern is that, for poorer children, college enrollment based on claims is much too high to represent actual college enrollment. While 1098T-based enrollment matches enrollment in the NLSY for children in the bottom quartile of family income, the claims-based measure of college enrollment is about twice this level. The most likely explanation for these implausibly-high claiming rates is that low-income families claim children after age 18 in order to claim the EITC[38]. This seems especially likely since the only reason most low-income families file a 1040 at all is to claim EITC benefits[39]. Therefore the claims-based measure of college enrollment cannot help us validate the 1098T-based measure of college enrollment for low-income children: false claims overwhelm genuine non-1098T college enrollment with zero tuition payments or at non-Title IV schools.

I would like to see if the main results in the paper hold up when using this alternative enrollment measure. For this to make sense, I restrict to richer families for two reasons. The first reason, as just discussed, is that this measure appears to approximate enrollment only for richer families. The second reason is that layoffs reduce filing rates, because filing correlates positively with income. This introduces a potential spurious effect by reducing claims by reducing filings. However, note that families who do not file are unlikely to have children in college, because such families can claim EITC benefits. Nonetheless, restricting to higher-income families alleviates most of the filing problem. I use the same definition of "high-income" used in the main results in the text, three-year mean incomes 1996-1999 above $40,000.

Figure 1.A5.2.a plots treatment effects from Equation (2) where the outcome is a dummy for whether the child is claimed by her parents, instead of whether the child receives a 1098T as in Figure 1.13. The claim-based enrollment measure shows almost exactly the same pattern of treatment effects as the 1098T-based measure for high-income families: about one percentage point

[38] Note that by using three-year mean family income on the x-axis, families with incomes in the EITC benefit range for at least one of these three years are likely to have much higher three-year mean income, due to mean-reversion, explaining why the excess claims extend "too high" relative to the EITC benefit range, which reaches zero around $30,000.

[39] Only 60% of the poorest parents in this graph file, where as almost 100% of parents file starting at around $50,000

lower enrollment for children experiencing a father's job loss one year before college decisions versus one year after college decisions. I conclude that this test supports the main results of the paper: layoffs reduce college enrollment by a small amount.

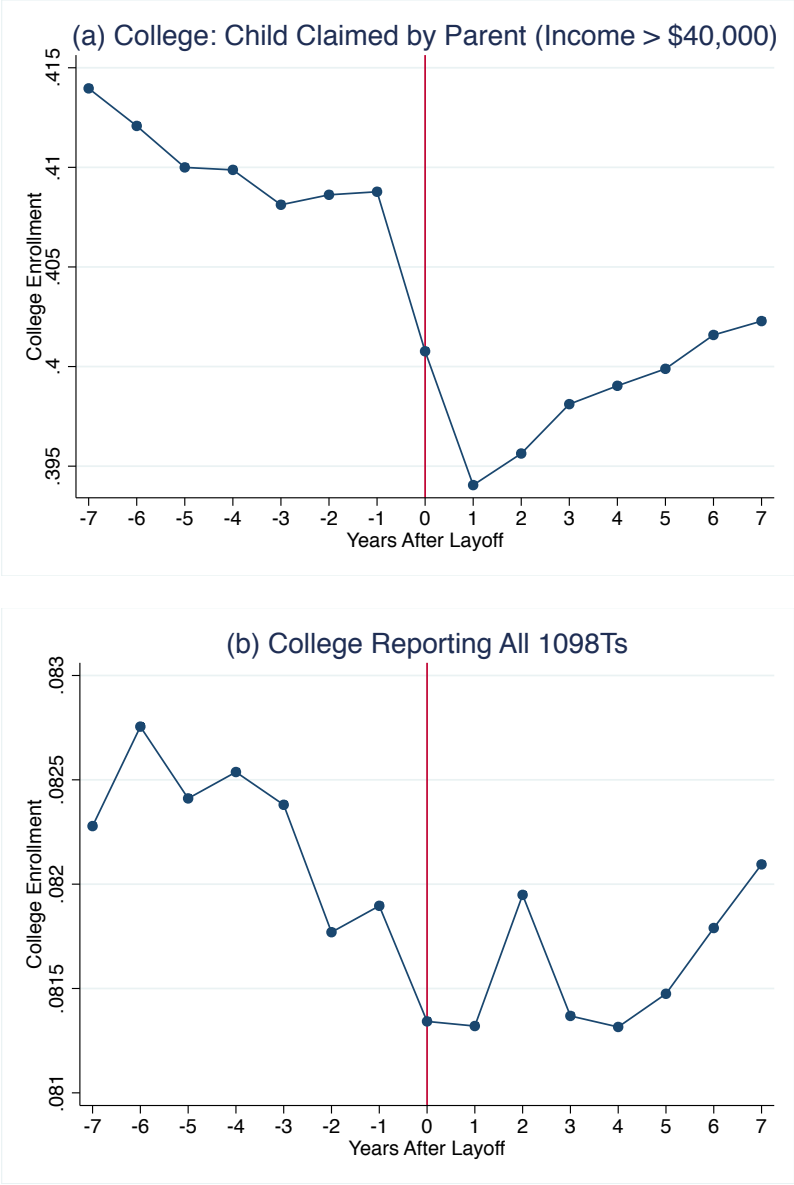


Figure 1.A5.2

Event-Age Studies for Alternative College Enrollment Measures

Notes: Panel (a) is analogous to Figure 1.4.b but replacing the 1098T indicator for college with a claimed-by-adult-filer indicator for college, and adding a restriction to age 19-22 and family incomes in 1996-1999 above $40,000. Panel (b) is also analogous to Figure 1.4.b but adding a restriction that the 1098T come from a college that appears to report 1098Ts even for children who make no payments for college.

I also develop a second robustness check on the main 1098T enrollment results. I create a dummy variable that equals one when a child receives a 1098T from a college that often files 1098Ts for students with zero net-tuition payments[40]. I then estimate Equation (2) again for this outcome variable, now for all children, not just high-income children. The resulting event-age effects are much noisier due to the much lower rate of enrollment at this restricted set of institutions, but the overall pattern is reassuring: children experiencing paternal layoffs one year before making enrollment decisions enroll in these colleges about 1% less than children experiencing paternal layoffs one year after making enrollment decisions. The only way this pattern could be generated spuriously is if these schools (1) raise financial aid for students who experience paternal layoff and then (2) selectively decide not to file 1098Ts for students whose greater aid package reduced their net payments to zero, despite filing 1098Ts for many other students with zero net payments. As this seems much less likely than the alternative explanation that paternal layoffs cause a small fraction of students to forego college, I conclude that this test also supports the main results of the paper.

[40] I define "often" by ranking colleges by the fraction of their 1098Ts recording zero net tuition payments, and restricting to those in the 75th percentile of this distribution.

References

1. Aaronson, Daniel, Kyung-Hong Park and Daniel Sullivan. 2007. "Explaining the Decline in Teen Labor Force Participation," Chicago Fed Letter 234.

2. Abowd, John M., Francis Kramarz and David N. Margolis. 1999. "High Wage Workers and High Wage Firms," *Econometrica* 67(2) 251-333

3. Acemoglu, Daron and J.S. Pischke. 2001. "Changes in the Wage Structure, Family Income, and Children's Education," *European Economic Review* 45: 890–904

4. Athreya, Kartik B., Devin Reilly and Nicole B. Simpson. 2010. "Earned Income Tax Credit Recipients: Income, Marginal Tax Rates, Wealth, and Credit Constraints," *Economic Quarterly* 96(3): 229-258

5. Attanasio, Orazio P. 1999. "Consumption." Handbook of Macroeconomics, ed. John B. Taylor and Michel Woodford, 741–812.

6. Bailey, Martha J. and Susan M. Dynarski. 2011. "Gains and Gaps: Changing Inequality in U.S. College Entry and Completion." NBER Working Paper 17633.

7. Becker, Gary S. 1994. Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education. University of Chicago Press.

8. Belley, Philippe and Lance Lochner. 2007. "The Changing Role of Family Income and Ability in Determining Educational Achievement." NBER Working Paper 13527.

9. Berndt, Ernst R., Zvi Griliches and Neal J. Rappaport. 1995. "Econometric Estimates of Price Indexes for Personal Computers in the 1990's." *Journal of Econometrics* 68: 243-268

10. Bettinger, Eric P., Bridget Terry Long, Philip Oreopoulos and Lisa Sanbonmatsu. 2009. "The Role of Simplification and Information in College Decisions: Results from the H&R Block FAFSA Experiment." NBER Working Paper 15361.

11. Bratberg, Espen, Oivind Anti Nilson, and Kjell Vaage. 2008. "Job Losses and Child Outcomes." *Labour Economics* 15: 591–603.

12. Bricker, Jesse, Arthur B. Kennickell, Kevin B Moore and John Sabelhaus. 2012. "Changes in U.S. Family Finances from 2007 to 2010: Evidence from the Survey of Consumer Finances." *Federal Reserve Bulletin* 98(2).

13. Carneiro, Pedro and James Heckman. 2002. "The Evidence on Credit Constraints in Post-Secondary Schooling," *The Economic Journal* 112: 705–734.

14. Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4): 1593-1660.

15. Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2011. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." NBER Working Paper 17699.

16. Chetty, Raj, John N. Friedman and Emmanuel Saez. 2012. "Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings." NBER Working Paper 18232.

17. Chetty, Raj and Adam Szeidl. 2007. "Consumption Commitments and Risk Preferences." *Quarterly Journal of Economics* 122(2): 831-877.

18. Coelli, Michael B. 2011. "Parental Job Loss and the Education Enrollment of Youth," *Labour Economics* 18: 25-35.

19. Currie, Janet. 2006. "The Take-up of Social Benefits." Poverty, the Distribution of Income, and Public Policy. ed. Alan Auerbach, David Card, and John Quigley. New York: Russell Sage.

20. Dahl, Gordon B. and Lance Lochner. 2012. "The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit." *The American Economic Review* 102(5): 1927-1956.

21. Delisle, Jason. 2012. "Federal Student Loan Interest Rates: History, Subsidies, and Cost." New America Foundation Issue Brief.

22. Deming, David and Susan Dynarski. 2010. "College Aid." Chapter 10 in Targeting Investments in Children: Fighting Poverty When Resources are Limited, ed. Phillip B. Levine and David J. Zimmerman. University of Chicago Press.

23. Edmonds, Eric V. 2007. "Child Labor." IZA Discussion Paper no. 2606.

24. Goldin, Claudia Dale and Lawrence F. Katz. 2008. The Race Between Education and Technology. Harvard University Press.

25. Grogger, Jeffrey. 1995. "The Effect of Arrests on the Employment and Earnings of Young Men." *Quarterly Journal of Economics* 110(1): 51-71.

26. Gruber, Jonathan. 1997. "The Consumption Smoothing Benefits of Unemployment Insurance." *The American Economic Review* 87(1): 192-205.

27. Hall, Bronwyn H., Jacques Mairesse, and Laure Turner. 2005. "Identifying Age, Cohort and Period Effects in Scientific Research Productivity: Discussion and Illustration Using Simulated and Actual Data on French Physicists," NBER Working Paper 11739.

28. Hoynes, Hilary W., Diane Whitmore Schanzenbach and Douglas Almond. 2012. "Long Run Impacts of Childhood Access to the Safety Net." Working Paper.

29. Hurst, Erik and Annamaria Lusardi. 2004. "Liquidity Constraints, Household Wealth, and Entrepreneurship." *The Journal of Political Economy* 112(2): 319-347.

30. Isaacs, Julia, Heather Hahn, Stephanie Rennane, C. Eugene Steuerle, Tracy Vericker. 2011. "Kids' Share: Report on Federal Expenditures on Children Through 2010." Urban Institute and Brookings Institution.

31. Jacobson, Louis S., Robert J. LaLonde and Daniel G. Sullivan. 1993. "Earnings Losses of Displaced Workers," *The American Economic Review*, 83(4): 685-709.

32. Keane, Michael P. and Kenneth I. Wolpin. 2001. "The Effect of Parental Transfers and Borrowing Constraints on Educational Attainment," *International Economic Review* Vol. 42(4).

33. Lazear, Edward. 1977. "Education: Consumption or Production?" *Journal of Political Economy* 85(3): 569-598.

34. Leslie, Larry L. 1984. "Changing Patterns in Student Financing of Higher Education." *Journal of Higher Education* 55(3): 313-346.

35. Loken, Katrine V, Magne Mogstad and Matthew Wiswall. 2012. "What Linear Estimators Miss: The Effects of Family Income on Child Outcomes." *American Economic Journal: Applied Economics* 4(2): 1-35.

36. Lovenheim, Michael F. 2011. "The Effect of Liquid Housing Wealth on College Enrollment." *Journal of Labor Economics* 29(4): 741-771.

37. Lovenheim, Michael F. and C. Lockwood Reynolds. 2012. "The Effect of Housing Wealth on College Choice: Evidence from the Housing Boom." NBER Working Paper 18075.

38. Mayer, Susan E. 1997. What Money Can't Buy: Family Income and Children's Life Chances. Harvard University Press.

39. Mayer, Susan E. 2010. "Revisiting an Old Question: How Much Does Parental Income Affect Child Outcomes?" *Focus* 27(2): 21-26.

40. Milligan, Kevin and Mark Stabile. 2008. "Do Child Tax Benefits Affect the Wellbing of Children? Evidence from Canadian Child Benefit Expansions," NBER Working Paper 14624.

41. Mueller, Andreas I. 2012. "Separations, Sorting and Cyclical Unemployment." IZA Discussion Paper 6849.

42. National Center for Education Statistics. 2006. "Student Financing of Undergraduate Education: 2003-04," Statistical Analysis Report 2006-185.

43. Oreopoulos, Philip, Marianne Page, and Ann Huff Stevens. 2008. "The Intergenerational Effects of Worker Displacement," *Journal of Labor Economics* 26(3): 455-483.

44. Paulin, Geoffrey D. 2001. "Expenditures of College-Age Students and Non-Students." *Monthly Labor Review*, July.

45. Rege, Mari, Kjetil Telle and Mark Votruba. 2011. "Parental Job Loss and Children's School Performance." *Review of Economic Studies* 78 (4): 1462-1489.

46. Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay and Student Achievement." *Quarterly Journal of Economics* 125 (1): 175-214.

47. Sallie Mae. 2011. "How America Pays for College 2011."

48. Shea, John. 2000. "Does Parents' Money Matter?" *Journal of Public Economics* 77: 155-184.

49. Stephens, Jr. Melvin. 2001. "The Long-Run Consumption Effects of Earnings Shocks," *The Review of Economics and Statistics* 83(1): 28-36.

50. ——— 2004. "Job Loss Expectations, Realizations, and Household Consumption Behavior." *The Review of Economics and Statistics* 86(1): 253-269.

51. Wachter, Til von and Daniel Sullivan. 2009. "Job Displacement and Mortality: An Analysis Using Administrative Data," *Quarterly Journal of Economics* 124 (3): 1265-1306.

52. Wachter, Til von, Jae Song and Joyce Manchester. 2009. "Long-Term Earnings Losses due to Mass Layoffs During the 1982 Recession: An Analysis Using U.S. Administrative Data from 1974 to 2004." Working Paper.

53. Weinberg, Bruce A. 2001. "An Incentive Model of the Effect of Parental Income on Children." *The Journal of Political Economy* 109(2): 266-280.

54. Yagan, Danny. 2012. "The 2003 Dividend Tax Cut and the Real Economy: Quasi-Experimental Evidence on Corporate Investment." Working Paper.

55. Zeldes, Stephen P. 1989. "Consumption and liquidity constraints: An empirical investigation." *The Journal of Political Economy* 97(2): 305-346.

II How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR[41]

II.A Introduction

What are the long-term impacts of early childhood education? Evidence on this important policy question remains scarce because of a lack of data linking childhood education and outcomes in adulthood. This paper analyzes the long-term impacts of Project STAR, one of the most widely studied education experiments in the United States. The Student/Teacher Achievement Ratio (STAR) experiment randomly assigned one cohort of 11,571 students and their teachers to different classrooms within their schools in grades K-3. Some students were assigned to small classes (15 students on average) in grades K-3, while others were assigned to large classes (22 students on average). The experiment was implemented across 79 schools in Tennessee from 1985 to 1989. Numerous studies have used the STAR experiment to show that class size, teacher quality, and peers have significant causal impacts on test scores (see Schanzenbach 2006 for a review). Whether these gains in achievement on standardized tests translate into improvements in adult outcomes such as earnings remains an open question.

We link the original STAR data to administrative data from tax returns, allowing us to follow 95% of the STAR participants into adulthood.[42] We use these data to analyze the impacts of STAR on outcomes ranging from college attendance and earnings to retirement savings, home ownership, and marriage. We begin by documenting the strong correlation between kindergarten test scores and adult outcomes. A one percentile increase in end-of-kindergarten (KG) test scores is associated with a $132 increase in wage earnings at age 27 in the raw data, and a $94 increase after controlling for parental characteristics. Several other adult outcomes – such as college attendance rates,

[41] This chapter is coauthored with Raj Chetty, John N. Friedman, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. We thank Lisa Barrow, David Card, Gary Chamberlain, Elizabeth Cascio, Janet Currie, Jeremy Finn, Edward Glaeser, Bryan Graham, James Heckman, Caroline Hoxby, Guido Imbens, Thomas Kane, Lawrence Katz, Alan Krueger, Derek Neal, Jonah Rockoff, Douglas Staiger, numerous seminar participants, and anonymous referees for helpful discussions and comments. We thank Helen Bain and Jayne Zaharias at HEROS for access to the Project STAR data. The tax data were accessed through contract TIRNO-09-R-00007 with the Statistics of Income (SOI) Division at the US Internal Revenue Service. Gregory Bruich, Jane Choi, Jessica Laird, Keli Liu, Laszlo Sandor, and Patrick Turley provided outstanding research assistance. Financial support from the Lab for Economic Applications and Policy at Harvard, the Center for Equitable Growth at UC Berkeley, and the National Science Foundation is gratefully acknowledged.

[42] The data for this project were analyzed through a program developed by the Statistics of Income (SOI) Division at the U.S. Internal Revenue Service to support research into the effects of tax policy on economic and social outcomes and improve the administration of the tax system.

quality of college attended, home ownership, and 401(k) savings – are also all highly correlated with kindergarten test scores. These strong correlations motivate the main question of the paper: do classroom environments that raise test scores – such as smaller classes and better teachers – cause analogous improvements in adult outcomes?

Our analysis of the experimental impacts combines two empirical strategies. First, we study the impacts of observable classroom characteristics. We analyze the impacts of class size using the same intent-to-treat specifications as Krueger (1999), who showed that students in small classes scored higher on standardized tests. We find that students assigned to small classes are 1.8 percentage points more likely to be enrolled in college at age 20, a significant improvement relative to the mean college attendance rate of 26.4% at age 20 in the sample. We do not find significant differences in earnings at age 27 between students who were in small and large classes, although these earnings impacts are imprecisely estimated. Students in small classes also exhibit statistically significant improvements on a summary index of the other outcomes we examine (home ownership, 401(k) savings, mobility rates, percent college graduate in ZIP code, and marital status).

We study variation across classrooms along other observable dimensions, such as teacher and peer characteristics, using a similar approach. Prior studies (e.g. Krueger 1999) have shown that STAR students with more experienced teachers score higher on tests. We find similar impacts on earnings. Students randomly assigned to a KG teacher with more than 10 years of experience earn an extra $\$1,093$ (6.9% of mean income) on average at age 27 relative to students with less experienced teachers.[43] We also test whether observable peer characteristics have long-term impacts by regressing earnings on the fraction of low-income, female, and black peers in KG. These peer impacts are not significant, but are very imprecisely estimated because of the limited variation in peer characteristics across classrooms.

Because we have few measures of observable classroom characteristics, we turn to a second empirical strategy that captures both observed and unobserved aspects of classrooms. We use an analysis of variance approach analogous to that in the teacher effects literature to test whether earnings are clustered by kindergarten classroom. Because we observe each teacher only once in our data, we can only estimate "class effects" – the combined effect of teachers, peers, and

[43] Because teacher experience is correlated with many other unobserved attributes – such as attachment to the teaching profession – we cannot conclude that increasing teacher experience would improve student outcomes. This evidence simply establishes that a student's KG teacher has effects on his or her earnings as an adult.

any class-level shock – by exploiting random assignment to KG classrooms of both students and teachers. Intuitively, we test whether earnings vary across KG classes by more than what would be predicted by random variation in student abilities. An F test rejects the null hypothesis that KG classroom assignment has no effect on earnings. The standard deviation of class effects on annual earnings is approximately 10% of mean earnings, highlighting the large stakes at play in early childhood education.

The analysis of variance shows that kindergarten classroom assignment has significant impacts on earnings, but it does not tell us whether classrooms that improve scores also generate earnings gains. That is, are class effects on earnings correlated with class effects on scores? To analyze this question, we proxy for each student's KG "class quality" by the average test scores of his classmates at the end of kindergarten. We show that end-of-class peer test scores are an omnibus measure of class quality because they capture peer effects, teacher effects, and all other classroom characteristics that affect test scores. Using this measure, we find that kindergarten class quality has significant impacts on both test scores and earnings. Students randomly assigned to a classroom that is one standard deviation higher in quality earn 3% more at age 27. Students assigned to higher quality classes are also significantly more likely to attend college, enroll in higher quality colleges, and exhibit improvements in the summary index of other outcomes. The class quality impacts are similar for students who entered the experiment in grades 1-3 and were randomized into classes at that point. Hence, the findings of this paper should be viewed as evidence on the long-term impacts of early childhood education rather than kindergarten in particular.

Our analysis of "class quality" must be interpreted very carefully. The purpose of this analysis is to detect clustering in outcomes at the classroom level: are a child's outcomes correlated with his peers' outcomes? Although we test for such clustering by regressing own scores and earnings on peer test scores, we emphasize that such regressions are *not* intended to detect peer effects. Because we use post-intervention peer scores as the regressor, these scores incorporate the impacts of peer quality, teacher quality, and any random class-level shock (such as noise from construction outside the classroom). The correlation between own outcomes and peer scores could be due to any of these factors. Our analysis shows that the classroom a student was assigned to in early childhood matters for outcomes 20 years later, but does not shed light on which specific factors should be manipulated to improve adult outcomes. Further research on which factors contribute

to high "class quality" would be extremely valuable in light of the results reported here.

The impacts of early childhood class assignment on adult outcomes may be particularly surprising because the impacts on test scores "fade out" rapidly. The impacts of class size on test scores become statistically insignificant by grade 8 (Krueger and Whitmore 2001), as do the impacts of class quality on test scores. Why do the impacts of early childhood education fade out on test scores but re-emerge in adulthood? We find some suggestive evidence that part of the explanation may be non-cognitive skills. We find that KG class quality has significant impacts on non-cognitive measures in 4th and 8th grade such as effort, initiative, and lack of disruptive behavior. These non-cognitive measures are highly correlated with earnings even conditional on test scores but are not significant predictors of future standardized test scores. These results suggest that high quality KG classrooms may build non-cognitive skills that have returns in the labor market but do not improve performance on standardized tests. While this evidence is far from conclusive, it highlights the value of further empirical research on non-cognitive skills.

In addition to the extensive literature on the impacts of STAR on test scores, our study builds on and contributes to a recent literature investigating selected long-term impacts of class size in the STAR experiment. These studies have shown that students assigned to small classes are more likely to complete high school (Finn, Gerber, and Boyd-Zaharias 2005) and take the SAT or ACT college entrance exams (Krueger and Whitmore 2001) and are less likely to be arrested for crime (Krueger and Whitmore 2001). Most recently, Muennig et al. (2010) report that students in small classes have higher mortality rates, a finding that we do not obtain in our data as we discuss below. We contribute to this literature by providing a unified evaluation of several outcomes, including the first analysis of earnings, and by examining the impacts of teachers, peers, and other attributes of the classroom in addition to class size.

Our results also complement the findings of studies on the long-term impacts of other early childhood interventions, such as the Perry and Abecederian preschool demonstrations and the Head Start program, which also find lasting impacts on adult outcomes despite fade-out on test scores (see Almond and Currie 2010 for a review). We show that a better classroom environment from ages 5-8 can have substantial long-term benefits even without intervention at earlier ages.

The paper is organized as follows. In Section II, we review the STAR experimental design and address potential threats to the validity of the experiment. Section III documents the cross-

sectional correlation between test scores and adult outcomes. Section IV analyzes the impacts of observable characteristics of classrooms – size, teacher characteristics, and peer characteristics – on adult outcomes. In Section V, we study class effects more broadly, incorporating unobservable aspects of class quality. Section VI documents the fade-out and re-emergence effects and the potential role of non-cognitive skills in explaining this pattern. Section VI concludes.

II.B Experimental Design and Data

II.B.1 Background on Project STAR

Word et al. (1990), Krueger (1999), and Finn et al. (2007) provide a comprehensive summary of Project STAR; here, we briefly review the features of the STAR experiment most relevant for our analysis. The STAR experiment was conducted at 79 schools across the state of Tennessee over four years. The program oversampled lower-income schools, and thus the STAR sample exhibits lower socioeconomic characteristics than the state of Tennessee and the U.S. population as a whole.

In the 1985-86 school year, 6,323 kindergarten students in participating schools were randomly assigned to a small (target size 13-17 students) or regular-sized (20-25 students) class within their schools.[44] Students were intended to remain in the same class type (small vs. large) through 3rd grade, at which point all students would return to regular classes for 4th grade and subsequent years. As the initial cohort of kindergarten students advanced across grade levels, there was substantial attrition because students who moved away from a participating school or were retained in grade no longer received treatment. In addition, because kindergarten was not mandatory and due to normal residential mobility, many children joined the initial cohort at the participating schools after kindergarten. A total of 5,248 students entered the participating schools in grades 1-3. These new entrants were randomly assigned to classrooms within school upon entry. Thus all students were randomized to classrooms within school upon entry, regardless of the entry grade. As a result, the randomization pool is school-by-entry-grade, and we include school-by-entry-grade fixed effects in all experimental analyzes below.

Upon entry into one of the 79 schools, the study design randomly assigned students not only

[44] There was also a third treatment group: regular sized class with a full-time teacher's aide. This was a relatively minor intervention, since all regular classes were already assigned a 1/3 time teacher's aide. Prior studies of STAR find no impact of a full-time teacher's aide on test scores. We follow the convention in the literature and group the regular and regular plus aide class treatments together.

to class type (small vs. large) but also to a classroom within each type (if there were multiple classrooms per type, as was the case in 50 of the 79 schools). Teachers were also randomly assigned to classrooms. Unfortunately, the exact protocol of randomization into specific classrooms was not clearly documented in any of the official STAR reports, where the emphasis was instead the random assignment into class type rather than classroom (Word et al. 1990). We present statistical evidence confirming that both students and teachers indeed appear to be randomly assigned directly to classrooms upon entry into the STAR project, as the original designers attest.

As in any field experiment, there were some deviations from the experimental protocol. In particular, some students moved from large to small classes and vice versa. To account for such potentially non-random sorting, we adopt the standard approach taken in the literature and assign treatment status based on initial random assignment (intent-to-treat).

In each year, students were administered the grade-appropriate Stanford Achievement Test, a multiple choice test that measures performance in math and reading. These tests were given only to students participating in STAR, as the regular statewide testing program did not extend to the early grades.[45] Following Krueger (1999), we standardize the math and reading scale scores in each grade by computing the scale score's corresponding percentile rank in the distribution for students in large classes. We then assign the appropriate percentile rank to students in small classes and take the average across math and reading percentile ranks. Note that this percentile measure is a ranking of students *within* the STAR sample.

II.B.2 Variable Definitions and Summary Statistics

We measure adult outcomes of Project STAR participants using administrative data from United States tax records. 95.0% of STAR records were linked to the tax data using an algorithm based on standard identifiers (SSN, date of birth, gender, and names) that is described in Online Appendix A.[46]

We obtain data on students and their parents from federal tax forms such as 1040 individual

[45] These K-3 test scores contain considerable predictive content. As reported in Krueger Whitmore (2001), the correlation between test scores in grades g and g+1 is 0.65 for KG and 0.80 for each grade 1-3. The values for grades 4-7 lie between 0.83 and 0.88, suggesting that the K-3 test scores contain similar predictive content.

[46] All appendix material is available as an on-line appendix posted as supplementary material to the article. Note that the matching algorithm was sufficiently precise that it uncovered 28 cases in the original STAR dataset that were a single split observation or duplicate records. After consolidating these records, we are left with 11,571 students.

income tax returns. Information from 1040's is available from 1996-2008. Approximately 10% of adults do not file individual income tax returns in a given year. We use third-party reports to obtain information such as wage earnings (form W-2) and college attendance (form 1098-T) for all individuals, including those who do not file 1040s. Data from these third-party reports are available since 1999. The year always refers to the tax year (i.e., the calendar year in which the income is earned or the college expense incurred). In most cases, tax returns for tax year t are filed during the calendar year $t+1$. The analysis dataset combines selected variables from individual tax returns, third party reports, and information from the STAR database, with individual identifiers removed to protect confidentiality.

We now describe how each of the adult outcome measures and control variables used in the empirical analysis is constructed. Table 2.1 reports summary statistics for these variables for the STAR sample as well as a random 0.25% sample of the US population born in the same years (1979-1980).

Table 2.1

SUMMARY STATISTICS

Variable	(1) STAR sample Mean	(2) STAR sample Std. Dev.	(3) U.S. 1979–80 cohort Mean	(4) U.S. 1979–80 cohort Std. Dev.
Adult outcomes				
Average wage earnings (2005–2007)	$15,912	$15,558	$20,500	$19,541
Zero wage earnings (2005–2007) (%)	13.9	34.5	15.6	36.3
Attended college in 2000 (age 20) (%)	26.4	44.1	34.7	47.6
College quality in 2000	$27,115	$4,337	$29,070	$7,252
Attended college by age 27 (%)	45.5	49.8	57.1	49.5
Owned a house by age 27 (%)	30.8	46.2	28.4	45.1
Made 401(k) contribution by age 27 (%)	28.2	45.0	31.0	46.2
Married by age 27 (%)	43.2	49.5	39.8	48.9
Moved out of TN by age 27 (%)	27.5	44.7		
Percent college graduates in 2007 ZIP code (%)	17.6	11.7	24.2	15.1
Deceased before 2010 (%)	1.70	12.9	1.02	10.1

Table 2.1 (Continued)

Variable	(1) STAR sample Mean	(2) STAR sample Std.Dev.	(3) U.S. 1979–80 cohort Mean	(4) U.S. 1979–80 cohort Std. Dev.
Parent characteristics				
Average household income (1996–98)	$48,014	$41,622	$65,661	$53,844
Mother's age at child's birth (years)	25.0	6.53	26.3	6.17
Married between 1996 and 2008 (%)	64.8	47.8	75.7	42.9
Owned a house between 1996 and 2008 (%)	64.5	47.8	53.7	49.9
Made a 401(k) contribution between 1996 and 2008 (%)	45.9	49.8	50.5	50.0
Missing (no parent found) (%)	13.9	34.6	23.9	42.6
Student background variables				
Female (%)	47.2	49.9	48.7	50.0
Black (%)	35.9	48.0		
Eligible for free or reduced-price lunch (%)	60.3	48.9		
Age at kindergarten entry (years)	5.65	0.56		
Teacher characteristics (entry-grade)				
Experience (years)	10.8	7.7		
Post-BA degree (%)	36.1	48.0		
Black (%)	19.5	39.6		
Number of observations	10,992		22,568	

Notes. Adult outcomes, parent characteristics, and student age at KG entry are from 1996–2008 tax data; other student background variables and teacher characteristics are from STAR database. Columns (1) and (2) are based on the sample of STAR students who were successfully linked to U.S. tax data. Columns (3) and (4) are based on a 0.25% random sample of the U.S. population born in the same years as the STAR cohort (1979–80). All available variables are defined identically in the STAR and U.S. samples. Earnings are average individual earnings in years 2005–2007, measured by wage earnings on W-2 forms; those with no W-2 earnings are coded as 0s. College attendance is measured by receipt of a 1098-T form, issued by higher education institutions to report tuition payments or scholarships. College quality is defined as the mean earnings of all former attendees of each college in the U.S. population at age 28. For individuals who did not attend college, college quality is defined by mean earnings at age 28 of those who did not attend college in the U.S. population. Home ownership is measured as those who report mortgage interest payments on a 1040 or 1098 tax form. 401(k) contributions are reported on W-2 forms. Marital status is measured by whether an individual files a joint tax return. State and ZIP code of residence are taken from the most recent 1040 form or W-2 form. Percent college graduates in the student's 2007 ZIP code is based on data on percent college graduates by ZIP code from the 2000 Census. Birth and death information are as recorded by the Social Security Administration. We link STAR participants to their parents by finding the earliest 1040 form in years 1996–2008 on which the STAR student is claimed as a dependent. We are unable to link 13.9% of the STAR children (and 23.9% of the U.S. cohort) to their parents; the summary statistics reported for parents exclude these observations. Parent household income is average adjusted gross income (AGI) in years 1996–1998, when STAR participants are aged 16–18. For years in which parents did not file, household income is defined as 0. For joint-filing parents, mother's age at child's birth uses the birth date of the female parent; for single-filing parents, the variable uses the birth date of the single parent, who is usually female. Other parent variables are defined in the same manner as student variables. Free or reduced-price lunch eligibility is an indicator for whether the student was ever eligible during the experiment. Student's age at kindergarten entry is defined as age (in days, divided by 365.25) on Sept. 1, 1985. Teacher experience is the number of years taught at any school before the student's year of entry into a STAR school. All monetary values are expressed in real 2009 dollars.

Earnings. The individual earnings data come from W-2 forms, yielding information on earnings for both filers and non-filers.[47] We define earnings in each year as the sum of earnings on all W-2 forms filed on an individual's behalf. We express all monetary variables in 2009 dollars, adjusting for inflation using the Consumer Price Index. We cap earnings in each year at $100,000 to reduce the influence of outliers; fewer than 1% of individuals in the STAR sample report earnings above $100,000 in a given year. To increase precision, we typically use average (inflation indexed) earnings from year 2005 to 2007 as an outcome measure. The mean individual earnings for the STAR sample in 2005-2007 (when the STAR students are 25-27 years old) is $15,912. This earnings measure includes zeros for the 13.9% of STAR students who report no income 2005-2007. The mean level of earnings in the STAR sample is lower than in the same cohort in the U.S. population, as expected given that Project STAR targeted more disadvantaged schools.

College Attendance. Higher education institutions eligible for federal financial aid – Title IV institutions – are required to file 1098-T forms that report tuition payments or scholarships received for every student.[48] Title IV institutions include all colleges and universities as well as vocational schools and other postsecondary institutions. Comparisons to other data sources indicate that 1098-T forms accurately capture US college enrollment.[49] We have data on college attendance from 1098-T forms for all students in our sample since 1999, when the STAR students were 19 years old. We define college attendance as an indicator for having one or more 1098-T forms filed on one's behalf in a given year. In the STAR sample, 26.4% of students are enrolled in college at age 20 (year 2000). 45.5% of students are enrolled in college at some point between 1999 and 2007, compared with 57.1% in the same cohort of the U.S. population. Because the data are based purely on tuition payments, we have no information about college completion or degree attainment.

College Quality. Using the institutional identifiers on the 1098-T forms, we construct an

[47] We obtain similar results using household adjusted gross income reported on individual tax returns. We focus on the W-2 measure because it provides a consistent definition of individual wage earnings for both filers and non-filers. One limitation of the W-2 easure is that it does not include self-employment income.

[48] These forms are used to administer the Hope and Lifetime Learning education tax credits created by the Taxpayer Relief Act of 1997. Colleges are not required to file 1098-T forms for students whose qualified tuition and related expenses are waived or paid entirely with scholarships or grants; however, in many instances the forms are available even for such cases, perhaps because of automation at the university level.

[49] In 2009, 27.4 million 1098-T forms were issued (Internal Revenue Service, 2010). According to the Current Population Survey (US Census Bureau, 2010, Tables V and VI), in October 2008, there were 22.6 million students in the U.S. (13.2 million full time, 5.4 million part-time, and 4 million vocational). As an individual can be a student at some point during the year but not in October and can receive a 1098-T form from more than one institution, the number of 1098-T forms for the calendar year should indeed be higher than the number of students as of October.

earnings-based index of college quality as follows. First, using the full population of all individuals in the United States aged 20 on 12/31/1999 and all 1098-T forms for year 1999, we group individuals by the higher education institution they attended in 1999. This sample contains over 1.4 million individuals.[50] We take a 1% sample of those not attending a higher education institution in 1999, comprising another 27,733 individuals, and pool them together in a separate "no college" category. Next, we compute average earnings of the students in 2007 when they are aged 28 by grouping students according to the educational institution they attended in 1999. This earnings-based index of college quality is highly correlated with the US News ranking of the best 125 colleges and universities: the correlation coefficient of our measure and the log US news rank is 0.75. The advantages of our index are that while the US News ranking only covers the top 125 institutions, ours covers all higher education institutions in the U.S. and provides a simple cardinal metric for college quality. Among colleges attended by STAR students, the average value of our earnings index is $35,080 for four-year colleges and $26,920 for two-year colleges.[51] For students who did not attend college, the imputed mean wage is $16,475.

Other Outcomes. We identify spouses using information from 1040 forms. For individuals who file tax returns, we define an indicator for marriage based on whether the tax return is filed jointly. We code non-filers as single because most non-filers in the U.S. who are not receiving Social Security benefits are single (Cilke 1998, Table I). We define a measure of ever being married by age 27 as an indicator for ever filing a joint tax return in any year between 1999 and 2007. By this measure, 43.2% of individuals are married at some point before age 27.

We measure retirement savings using contributions to 401(k) accounts reported on W-2 forms from 1999-2007. 28.2% of individuals in the sample make a 401(k) contribution at some point during this period. We measure home ownership using data from the 1098 form, a third party report filed by lenders to report mortgage interest payments. We include the few individuals who report a mortgage deduction on their 1040 forms but do not have 1098's as homeowners. We define any individual who has a mortgage interest deduction at any point between 1999 and 2007 as a homeowner. Note that this measure of home ownership does not cover individuals who own homes

[50] Individuals who attended more than one institution in 1999 are counted as students at all institutions they attended.

[51] For the small fraction of STAR students who attend more than one college in a single year, we define college quality based on the college that received the largest tuition payments on behalf of the student.

without a mortgage, which is rare among individuals younger than 27. By our measure, 30.8% of individuals own a home by age 27. We use data from 1040 forms to identify each household's ZIP code of residence in each year. For non-filers, we use the ZIP code of the address to which the W-2 form was mailed. If an individual did not file and has no W-2 in a given year, we impute current ZIP code as the last observed ZIP code. We define a measure of cross-state mobility by an indicator for whether the individual ever lived outside Tennessee between 1999 and 2007. 27.5% of STAR students lived outside Tennessee at some point between age 19 and 27. We construct a measure of neighborhood quality using data on the percentage of college graduates in the individual's 2007 ZIP code from the 2000 Census. On average, STAR students lived in 2007 in neighborhoods with 17.6% college graduates.

We observe dates of birth and death until the end of 2009 as recorded by the Social Security Administration. We define each STAR participant's age at kindergarten entry as the student's age (in days divided by 365.25) as of September 1, 1985. Virtually all students in STAR were born in the years 1979-1980. To simplify the exposition, we say that the cohort of STAR children is aged a in year $1980 + a$ (e.g., STAR children are 27 in 2007). Approximately 1.7% of the STAR sample is deceased by 2009.

Parent Characteristics. We link STAR children to their parents by finding the earliest 1040 form from 1996-2008 on which the STAR student was claimed as dependents. Most matches were found on 1040 forms for the tax year 1996, when the STAR children were 16. We identify parents for 86% of the STAR students in our linked dataset. The remaining students are likely to have parents who did not file tax returns in the early years of the sample when they could have claimed their child as a dependent, making it impossible to link the children to their parents. Note that this definition of parents is based on who claims the child as a dependent, and thus may not reflect the biological parent of the child.

We define parental household income as average Adjusted Gross Income (capped at $252,000, the 99th percentile in our sample) from 1996-1998, when the children were 16-18 years old. For years in which parents did not file, we define parental household income as zero. For divorced parents, this income measure captures the total resources available to the household claiming the

[51] Alternative definitions of income for non-filers – such as income reported on W-2's starting in 1999 – yield very similar results to those reported below.

child as a dependent (including any alimony payments), rather than the sum of the individual incomes of the two parents. By this measure, mean parent income is $48,010 (in 2009 dollars) for STAR students whom we are able to link to parents. We define marital status, home ownership, and 401(k) saving as indicators for whether the parent who claims the STAR child ever files a joint tax return, has a mortgage interest payment, or makes a 401(k) contribution over the period for which relevant data are available. We define mother's age at child's birth using data from Social Security Administration records on birth dates for parents and children. For single parents, we define the mother's age at child's birth using the age of the filer who claimed the child, who is typically the mother but is sometimes the father or another relative.[52] By this measure, mothers are on average 25.0 years old when they give birth to a child in the STAR sample. When a child cannot be matched to a parent, we define all parental characteristics as zero, and we always include a dummy for missing parents in regressions that include parent characteristics.

Background Variables from STAR. In addition to classroom assignment and test score variables, we use some demographic information from the STAR database in our analysis. This includes gender, race (an indicator for being black), and whether the student ever received free or reduced price lunch during the experiment. 36% of the STAR sample are black and 60% are eligible for free or reduced-price lunches. Finally, we use data on teacher characteristics – experience, race, and highest degree – from the STAR database. The average student has a teacher with 10.8 years of experience. 19.5% of kindergarten students have a black teacher, and 35.9% have a teacher with a master's degree or higher.

Our analysis dataset contains one observation for each of the 10,992 STAR students we link to the tax data. Each observation contains information on the student's adult outcomes, parent characteristics, and classroom characteristics in the grade the student *entered* the STAR project and was randomly assigned to a classroom. Hence, when we pool students across grades, we include test score and classroom data only from the entry grade.

[52] We define the mother's age at child's birth as missing for 471 observations in which the implied mother's age at birth based on the claiming parent's date of birth is below 13 or above 65. These are typically cases where the parent does not have an accurate birth date recorded in the SSA file.

II.B.3 Validity of the Experimental Design

The validity of the causal inferences that follow rests on two assumptions: successful randomization of students into classrooms and no differences in attrition (match rates) across classrooms. We now evaluate each of these issues.

Randomization into Classrooms. To evaluate whether the randomization protocol was implemented as designed, we test for balance in pre-determined variables across classrooms. The original STAR dataset contains only a few pre-determined variables: age, gender, race, and free-lunch status. Although the data are balanced on these characteristics, some skepticism naturally has remained because of the coarseness of the variables (Hanushek 2003).

The tax data allow us to improve upon the prior evidence on the validity of randomization by investigating a wider variety of family background characteristics. In particular, we check for balance in the following five parental characteristics: household income, 401(k) savings, home ownership, marital status, and mother's age at child's birth. Although most of these characteristics are not measured prior to random assignment in 1985, they are measured prior to the STAR cohort's expected graduation from high school and are unlikely to be impacted by the child's classroom assignment in grades K-3. We first establish that these parental characteristics are in fact strong predictors of student outcomes. In column 1 of Table 2.2, we regress the child's earnings on the five parent characteristics, the student's age, gender, race, and free-lunch status, and school-by-entry-grade fixed effects. We also include indicators for missing data on certain variables (parents' characteristics, mother's age, student's free lunch status, and student's race). The student and parent demographic characteristics are highly significant predictors of earnings.

Table 2.2

RANDOMIZATION TESTS

Dependent variable	(1) Wage earnings (%)	(2) Small class (%)	(3) Teacher experience (years)	(4) Teacher has post-BA deg. (%)	(5) Teacher is Black (%)	(6) *p*-value
Parent's income ($1000s)	65.47	−0.003	−0.001	0.016	−0.003	0.848
	(6.634)	(0.015)	(0.002)	(0.012)	(0.007)	
	[9.87]	[−0.231]	[−0.509]	[1.265]	[−0.494]	
Mother's age at STAR birth	53.96	0.029	0.022	0.008	0.060	0.654
	(24.95)	(0.076)	(0.012)	(0.061)	(0.050)	
	[2.162]	[0.384]	[1.863]	[0.132]	[1.191]	
Parents have 401(k)	2,273	1.455	0.111	0.431	−1.398	0.501
	(348.3)	(1.063)	(0.146)	(0.917)	(0.736)	
	[6.526]	[1.368]	[0.761]	[0.469]	[−1.901]	
Parents own home	390.9	−0.007	−0.023	−2.817	0.347	0.435
	(308.1)	(0.946)	(0.159)	(0.933)	(0.598)	
	[1.269]	[−0.008]	[−0.144]	[−3.018]	[0.58]	
Parents married	968.3	0.803	0.166	−0.306	−0.120	0.820
	(384.2)	(1.077)	(0.165)	(1.101)	(0.852)	
	[2.52]	[0.746]	[1.008]	[−0.277]	[−0.14]	
Student female	−2,317	−0.226	0.236	−0.057	−0.523	0.502
	(425.0)	(0.864)	(0.111)	(0.782)	(0.521)	
	[−5.451]	[−0.261]	[2.129]	[−0.072]	[−1.003]	
Student black	−620.8	0.204	0.432	2.477	1.922	0.995
	(492.0)	(1.449)	(0.207)	(1.698)	(1.075)	
	[−1.262]	[0.141]	[2.089]	[1.459]	[1.788]	
Student free-lunch	−3,829	−0.291	0.051	−0.116	−0.461	0.350
	(346.2)	(1.110)	(0.149)	(0.969)	(0.648)	
	[−11.06]	[−0.262]	[0.344]	[−0.12]	[−0.712]	
Student's age at KG entry	−2,001	−0.828	−0.034	0.140	−0.364	0.567
	(281.4)	(0.885)	(0.131)	(0.738)	(0.633)	
	[−7.109]	[−0.935]	[−0.257]	[0.19]	[−0.575]	
Student predicted earnings						0.916
p-value of *F* test	0.000	0.261	0.190	0.258	0.133	
Observations	10,992	10,992	10,914	10,938	10,916	

Notes. Columns (1)–(5) each report estimates from an OLS regression of the dependent variable listed in the column on the variables listed in the rows and school-by-entry-grade fixed effects. The regressions include one observation per student, pooling across all entry grades. Standard errors clustered by school are reported in parentheses and *t*-statistics in square brackets. Small class is an indicator for assignment to a small class on entry. Teacher characteristics are for teachers in the entry grade. Independent variables are predetermined parent and student characteristics. See notes to TableI for definitions of these variables. The *p*-value reported at bottom of columns (1)–(5) is for an *F* test of the joint significance of the variables listed in the rows. Each row of column (6) reports a *p*-value from a separate OLS regression of the predetermined variable listed in the corresponding row on school and class fixed effects (omitting one class per school). The *p*-value is for an *F* test of the joint significance of the class fixed effects. The *F* tests in column (6) use the subsample of students who entered in kindergarten. Student predicted earnings is formed using the specification in column (1), excluding the school-by-entry-grade fixed effects. Some observations have missing data on parent characteristics, free-lunch status, race, or mother's age at STAR birth. Columns (1)–(5) include these observations along with four indicators for missing data on these variables. In column (6), observations with missing data are excluded from the regressions with the corresponding dependent variables.

Having identified a set of pre-determined characteristics that predict children's future earnings, we test for balance in these covariates across classrooms. We first evaluate randomization into the small class treatment by regressing an indicator for being assigned to a small class upon entry on the same variables as in column 1. As shown in column 2 of Table 2.2, none of the demographic characteristics predict the likelihood that a child is assigned to a small class. An F test for the joint significance of all the pre-determined demographic variables is insignificant ($p = 0.26$), showing that students in small and large classes have similar demographic characteristics.

Columns 3-5 of Table 2.2 evaluate the random assignment of teachers to classes by regressing teacher characteristics – experience, bachelor's degree, and race – on the same student and parent characteristics. Again, none of the pre-determined variables predict the type of teacher a student is assigned, consistent with random assignment of teachers to classrooms.

Finally, we evaluate whether students were randomly assigned into classrooms within small or large class types. If students were randomly assigned to classrooms, then conditional on school fixed effects, classroom indicator variables should not predict any pre-determined characteristics of the students. Column 6 of Table 2.2 reports p values from F tests for the significance of kindergarten classroom indicators in regressions of each pre-determined characteristic on class and school fixed effects. None of the F tests is significant, showing that each of the parental and child characteristics is balanced across classrooms. To test whether the pre-determined variables jointly predict classroom assignment, we predict earnings using the specification in column 1 of Table II. We then regress predicted earnings on KG classroom indicators and school fixed effects and run an F test for the significance of the classroom indicators. The p value of this F test is 0.92, confirming that one would not predict clustering of earnings by KG classroom based on pre-determined variables. We use only kindergarten entrants for the F tests in column 6 because F tests for class effects are not powerful in grades 1-3 as only a few students enter each class in those grades. In Online Appendix Table II, we extend these randomization tests to include students who entered in grades 1-3 using the technique developed in Section V below and show that covariates are balanced across classrooms in later entry grades as well.

Selective Attrition. Another threat to the experimental design is differential attrition across classrooms (Hanushek 2003). Attrition is a much less serious concern in the present study than in past evaluations of STAR because we are able to locate 95% of the students in the tax data.

Nevertheless, we investigate whether the likelihood of being matched to the tax data varies by classroom assignment within schools. In columns 1 and 2 of Table 2.3, we test whether the match rate varies significantly with class size by regressing an indicator for being matched on the small class dummy. Column 1 includes no controls other than school-by-entry-grade fixed effects. It shows that, eliminating the between-school variation, the match rate in small and large classes differs by less than 0.02 percentage points. Column 2 shows that controlling for the full set of demographic characteristics used in Table 2.2 does not uncover any significant difference in the match rate across class types. The p values reported at the bottom of columns 1 and 2 are for F tests of the significance of classroom indicators in predicting match rates in regression specifications analogous to those in column 6 of Table 2.2. The p values are approximately 0.9, showing that there are no significant differences in match rates across classrooms within schools.

Table 2.3

TESTS FOR DIFFERENTIAL MATCH AND DEATH RATES

Dependent variable	(1) Matched (%)	(2) Matched (%)	(3) Deceased (%)	(4) Deceased (%)
Small class	−0.019	0.079	−0.010	−0.006
	(0.467)	(0.407)	(0.286)	(0.286)
p-value on F test on class effects	0.951	0.888	0.388	0.382
Demographic controls		x		x
Mean of dep. var.	95.0	95.0	1.70	1.70

Notes. The first row of each column reports coefficients from OLS regressions on a small class indicator and school-by-entry-grade fixed effects, with standard errors clustered by school in parentheses. The second row reports a p-value from a separate OLS regression of the dependent variable on school and class fixed effects (omitting one class per school). The p-value is for an F test of the joint significance of the class fixed effects. Matched is an indicator for whether the STAR student was located in the tax data using the algorithm described in Appendix A. Deceased is an indicator for whether the student died before 2010 as recorded by the Social Security Administration. Columns (1)–(2) are estimated on the full sample of students in the STAR database; columns (3) and (4) are estimated on the sample of STAR students linked to the tax data. Specifications (2) and (4) control for the following demographic characteristics: student gender, free-lunch status, age, and race, and a quartic in the claiming parent's household income interacted with parent's marital status, mother's age at child's birth, whether the parents own a home, and whether the parents make a 401(k) contribution between 1996 and 2008. Some observations have missing data on parent characteristics, free-lunch status, race, and mother's age at STAR birth; these observations are included along with four indicators for missing data on these variables.

Another potential source of attrition from the sample is through death. Columns 3 and 4 replicate the first two columns, replacing the dependent variable in the regressions with an indicator for death before January 1, 2010. We find no evidence that mortality rates vary with class size or across classrooms. The difference in death rates between small and large classes is approximately 0.01 percentage points. This finding is inconsistent with recent results reported by Muennig et al. (2010), who find that students in small classes and regular classes with a certified teaching assistant are slightly more likely to die using data from the National Death Index. We find that 154 STAR students have died by 2007 while Muennig et al. (2010) find 141 deaths in their data. The discrepancy between the findings might be due to differences in match quality.[53]

II.C Test Scores and Adult Outcomes in the Cross-Section

We begin by documenting the correlations between test scores and adult outcomes in the cross-section to provide a benchmark for assessing the impacts of the randomized interventions. Figure 2.1a documents the association between end-of-kindergarten test scores and mean earnings from age 25-27.[54] To construct this figure, we bin individuals into twenty equal-width bins (vingtiles) and plot mean earnings in each bin. A one percentile point increase in KG test score is associated with a \$132 (0.83%) increase in earnings twenty years later. If one codes the x-axis using national percentiles on the standardized KG tests instead of within-sample percentiles, the earnings increase is \$154 per percentile. The correlation between KG test score percentiles and earnings is linear and remains significant even in the tails of the distribution of test scores. However, KG test scores explain only a small share of the variation in adult earnings: the adjusted R^2 of the regression of earnings on scores is 5%.[55]

[53] As 95% of STAR students are matched to the our data and have a valid Social Security Number, we believe that deaths are recorded accurately in our sample. It is unclear why a lower match rate would lead to a systematic difference in death rates by class size. However, given the small number of deaths, slight imbalances might generate marginally significant differences in death rates across class types.

[54] Although individuals' earnings trajectories remain quite steep at age 27, earnings levels from ages 25-27 are highly correlated with earnings at later ages (Haider and Solon 2006), a finding we have confirmed with our population wide longitudinal data (see Online Appendix Table I).

[55] These cross-sectional estimates are consistent with those obtained by Currie and Thomas (2001) using the British National Child Development Survey and Currie (2010) using the National Longitudinal Survey of Youth.

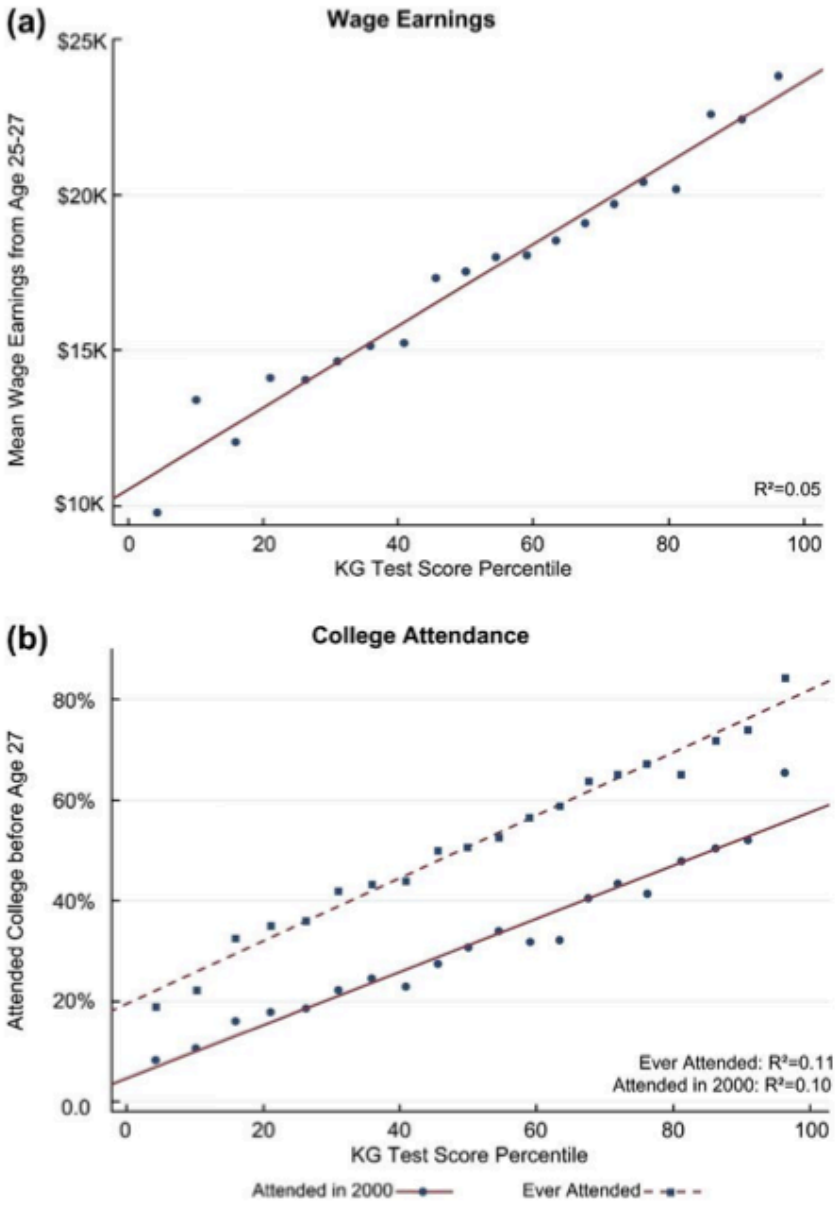


Figure 2.1

Correlation between Kindergarten Test Scores and Adult Outcomes

This figure plots the raw correlations between adult outcomes and kindergarten average test scores in math and reading (measured by within-sample percentile ranks). To construct these figures, we bin test scores into twenty equal sized (5 percentile point) bins and plot the mean of the adult outcome within each bin. The solid or dashed line shows the best linear fit estimated on the underlying student-level data using OLS. The R^2 from this regression, listed in each panel, shows how much of the variance in the outcome is explained by KG test scores. Earnings are mean annual earnings over years 2005-2007, measured by wage earnings on W-2 forms; those with no W-2 earnings are coded as zeros. College attendance is measured by receipt of a 1098-T form, issued by higher education

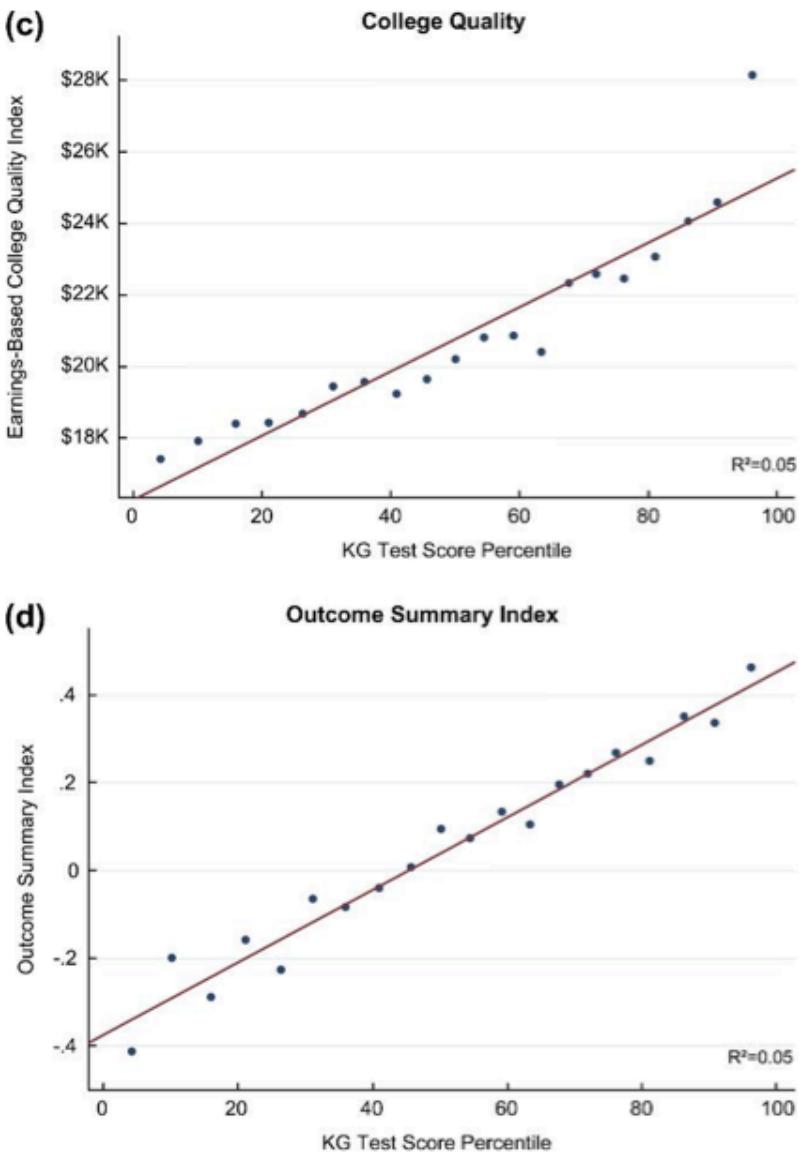


Figure 2.1 (Continued)

institutions to report tuition payments or scholarships, at some point between 1999 and 2007. The earnings-based index of college quality is a measure of the mean earnings of all former attendees of each college in the U.S. population at age 28, as described in the text. For individuals who did not attend college, college quality is defined by mean earnings at age 28 of those who did not attend college in the U.S. population. The summary index is the standardized sum of five measures, each standardized on its own before the sum: home ownership, 401(k) retirement savings, marital status, cross-state mobility, and percent of college graduates in the individual's 2007 ZIP code of residence. Thus the summary index has mean 0 and standard deviation of 1. All monetary values are expressed in real 2009 dollars.

Figures 2.1b and 2.1c show that KG test scores are highly predictive of college attendance rates and the quality of the college the student attends, as measured by our earnings-based index of college quality. To analyze the other adult outcomes in a compact manner, we construct a summary index of five outcomes: ever owning a home by 2007, 401(k) savings by 2007, ever married by 2007, ever living outside Tennessee by 2007, and living in a higher SES neighborhood in 2007 as measured by the percent of college graduates living in the ZIP code. Following Kling, Liebman, and Katz (2007), we first standardize each outcome by subtracting its mean and dividing it by its standard deviation. We then sum the five standardized outcomes and divide by the standard deviation of the sum to obtain an index that has a standard deviation of 1. A higher value of the index represents more desirable outcomes. Students with higher entry-year test scores have stronger adult outcomes as measured by the summary index, as shown in Figure 2.1d.

The summary index should be interpreted as a broader measure of success in young adulthood. Some of its elements proxy for future earnings conditional on current income. For example, having 401(k) savings reflects holding a good job that offers such benefits. Living outside Tennessee is a proxy for cross-state mobility, which is typically associated with higher socio-economic status. While none of these outcomes are unambiguously positive – for instance, marriage or homeownership by age 27 could in principle reflect imprudence – existing evidence suggests that, on net, these measures are associated with better outcomes. In our sample, each of the five outcomes is highly positively correlated with test scores on its own, as shown in Online Appendix Table III.

Table 2.4 quantifies the correlations between test scores and adult outcomes. We report standard errors clustered by school in this and all subsequent tables. Column 1 replicates Figure 2.1a by regressing earnings on KG test scores without any additional controls. Column 2 controls for classroom fixed effects and a vector of parent and student demographic characteristics. The parent characteristics are a quartic in parent's household income interacted with an indicator for whether the filing parent is ever married between 1996 and 2008, mother's age at child's birth, and indicators for parent's 401(k) savings and home ownership. The student characteristics are gender, race, age at entry-year entry, and free lunch status.[56] We use this vector of demographic characteristics in most specifications below. When the class fixed effects and demographic controls

[56] We code all parental characteristics as 0 for students whose parents are missing, and include an indicator for missing parents as a control. We also include indicators for missing data on certain variables (mother's age, student's free lunch status, and student's race) and code these variables as zero when missing.

are included, the coefficient on kindergarten percentile scores falls to \$94, showing that part of the raw correlation in Figure 2.1a is driven by these characteristics. Equivalently, a one standard deviation (SD) increase in test scores is associated with an 18% increase in earnings conditional on demographic characteristics.

Table 2.4

CROSS-SECTIONAL CORRELATION BETWEEN TEST SCORES AND ADULT OUTCOMES

Dependent variable	(1)	(2)	(3) Wage earnings ($)	(4)	(5)	(6) College in 2000 (%)	(7) College by age 27 (%)	(8) College quality (%)	(9) Summary index (%SD)
Entry-grade test percentile	131.7	93.79	90.04	0.102	97.7	0.364	0.510	32.04	0.551
	(12.24)	(11.63)	(8.65)	(12.87)	(8.47)	(0.022)	(0.021)	(3.40)	(0.048)
Eighth-grade test percentile				148.2					
				(11.95)					
Parental income percentile					145.5				
					(8.15)				
Entry grade	KG	KG	All	All	All	All	All	All	All
Class fixed effects		x	x	x	x	x	x	x	x
Student controls		x	x	x	x	x	x	x	x
Parent controls		x	x	x		x	x	x	x
Adjusted R^2	0.05	0.17	0.17	0.17	0.16	0.26	0.28	0.19	0.23
Observations	5,621	5,621	9,939	7,069	9,939	9,939	9,939	9,939	9,939

Notes. Each column reports coefficients from an OLS regression, with standard errors clustered by school in parentheses. In columns (1)–(2), the sample includes only kindergarten entrants; in columns (3)–(9), the sample includes all entry grades. Test percentile is the within-sample percentile rank of the student's average score in math and reading. Entry grade is the grade (kindergarten, 1, 2, or 3) when the student entered a STAR school. Entry-grade test percentile refers to the test score from the end of the student's first year at a STAR school. Grade 8 scores are available for students who remained in Tennessee public schools and took the eighth-grade standardized test any time between 1990 and 1997. Parental income percentile is the parent's percentile rank in the U.S. population household income distribution. Columns with class fixed effects isolate nonexperimental variation in test scores. Columns (2)–(9) all control for the following student characteristics: race, gender, and age at kindergarten. Parent controls comprise the following: a quartic in parent's household income interacted with an indicator for whether the filing parent is ever married between 1996 and 2008, mother's age at child's birth, indicators for parent's 401(k) savings and home ownership, and student's free-lunch status. The dependent variable in columns (1)–(5) is mean wage earnings over years 2005–2007 (including 0s for people with no wage earnings). College attendance is measured by receipt of a 1098-T form, issued by higher education institutions to report tuition payments or scholarships. The earnings-based index of college quality is a measure of the mean earnings of all former attendees of each college in the U.S. population at age 28, as described in the text. For individuals who did not attend college, college quality is defined by mean earnings at age 28 of those who did not attend college in the U.S. population. Summary index is the standardized sum of five measures, each standardized on its own before the sum: home ownership, 401(k) retirement savings, marital status, cross-state mobility, and percentage of college graduates in the individual's 2007 ZIP code of residence.

Columns 1 and 2 use only kindergarten entrants. 55% of students entered STAR in Kindergarten, with 20%, 14% and 11% entering in grades 1 through 3, respectively. In column 3, we also include students who entered in grades 1-3 in order to obtain estimates consistent with the experimental analysis below, which pools all entrants. To do so, we define a student's "entry-grade" test score as her score at the end of the grade in which she entered the experiment. Column 3 shows that a 1 percentile increase in entry-grade scores is associated with a $90 increase in earnings conditional on demographic controls. This $90 coefficient is a weighted average of the correlations between grade K-3 test scores and earnings, with the weights given by the entry rates in each grade.

In column 4, we include both 8th grade scores (the last point at which data from standardized tests are available for most students in the STAR sample) and entry-grade scores in the regression. The entire effect of entry-grade test score is absorbed by the 8th grade score, but the adjusted R^2 is essentially unchanged. In column 5, we compare the relative importance of parent characteristics and cognitive ability as measured by test scores. We calculate the parent's income percentile rank using the tax data for the U.S. population. We regress earnings on test scores, parents' income percentile, and controls for the student's race, gender, age, and class fixed effects. A one percentile point increase in parental income is associated with approximately a $148 increase in earnings, suggesting that parental background affects earnings as much as or more than cognitive ability in the cross section.[57]

Columns 6-9 of Table 2.4 show the correlations between entry-grade test scores and the other outcomes we study. Conditional on demographic characteristics, a one percentile point increase in entry-grade score is associated with a 0.36 percentage point increase in the probability of attending college at age 20 and a 0.51 percentage point increase in the probability of attending college at some point before age 27. A one percentile point increase in score is associated with $32 higher predicted earnings based on the college the student attends and a 0.5% of a standard deviation improvement in the summary index of other outcomes.

We report additional cross-sectional correlations in the online appendix. Online Appendix Table IV replicates Table 2.4 for each entry grade separately. Online Appendix Table V documents the correlation between test scores and earnings from grades K-8 for a fixed sample of students, while Online Appendix Table VI reports the heterogeneity of the correlations by race, gender, and

[57] Moreover, this $148 coefficient is an underestimate if parental income directly affects entry-grade test scores.

free lunch status. Throughout, we find very strong correlations between test scores and adult outcomes, which motivates the central question of the paper: do classroom environments that raise early childhood test scores also yield improvements in adult outcomes?

II.D Impacts of Observable Classroom Characteristics

In this section, we analyze the impacts of three features of classrooms that we can observe in our data – class size, teacher characteristics, and peer characteristics.

II.D.1 Class Size

We estimate the effects of class size on adult outcomes using an intent-to-treat regression specification analogous to Krueger (1999):

$$y_{icnw} = \alpha_{nw} + \beta \text{SMALL}_{cnw} + X_{icnw}\delta + \varepsilon_{icnw} \tag{10}$$

where y_{icnw} is an outcome such as earnings for student i randomly assigned to classroom c at school n in entry grade (wave) w. The variable $SMALL_{cnw}$ is an indicator for whether the student was assigned to a small class upon entry. Because children were randomly assigned to classrooms within schools in the first year they joined the STAR cohort, we include school-by-entry-grade fixed effects (α_{nw}) in all specifications. The vector X_{icnw} includes the student and parent demographic characteristics described above: a quartic in household income interacted with an indicator for whether the parents are ever married, 401(k) savings, home ownership, mother's age at child's birth, and the student's gender, race, age (in days), and free lunch status (along with indicators for missing data). To examine the robustness of our results, we report the coefficient both with and without this vector of controls. The inclusion of these controls does not significantly affect the estimates, as expected given that the covariates are balanced across classrooms. In all specifications, we cluster standard errors by school. Although treatment occurred at the classroom level, clustering by school provides a conservative estimate of standard errors that accounts for any cross-classroom correlations in errors within schools, including across students in different entry grades. These standard errors are in nearly all cases larger than those from clustering on only

classroom.[58]

We report estimates of equation (10) for various outcomes in Table V using the full sample of STAR students; we show in Online Appendix Table VIII that similar results are obtained for the subsample of students who entered in kindergarten. As a reference, in column 1 of Table 2.5, we estimate equation (10) with the entry grade test score as the outcome. Consistent with Krueger (1999), we find that students assigned to small classes score 4.8 percentile points higher on tests in the year they enter a participating school. Note that the average student assigned to a small class spent 2.27 years in a small class, while those assigned to a large class spent 0.13 years in a small class. On average, large classes had 22.6 students while small classes had 15.1 students. Hence, the impacts on adult outcomes below should be interpreted as effects of attending a class that is 33% smaller for 2.14 years.

[58] Online Appendix Table VII compares standard errors when clustering at different levels for key specifications.

Table 2.5

EFFECTS OF CLASS SIZE ON ADULT OUTCOMES

Dependent variable	(1) Test score (%)	(2) College in 2000 (%)	(3) College by age 27 (%)	(4) College quality ($)	(5) Wage earnings ($)	(6) Summary index (%of SD)
Small class (no controls)	4.81	2.02	1.91	119	4.09	5.06
	(1.05)	(1.10)	(1.19)	(96.8)	(327)	(2.16)
Small class (with controls)	4.76	1.78	1.57	109	−124	4.61
	(0.99)	(0.95)	(1.07)	(92.6)	(336)	(2.09)
Observations	9,939	10,992	10,992	10,992	10,992	10,992
Mean of dep. var.	48.67	26.44	45.50	27,115	15,912	0.00

Notes. Each column reports the coefficient on an indicator for initial small class assignment from two separate OLS regressions, with standard errors clustered by school in parentheses. All specifications include school-by-entry-grade fixed effects to isolate random variation in class assignment. The estimates in the second row (with controls) are from specifications that additionally control for the full vector of demographic characteristics used first in TableIV: a quartic in parent's household income interacted with an indicator for whether the filing parent is ever married between 1996 and 2008, mother's age at child's birth, indicators for parent's 401(k) savings and home ownership, and student's race, gender, free-lunch status, and age at kindergarten. Test score is the average math and reading percentile rank score attained in the student's year of entry into the experiment. Wage earnings are the mean earnings across years 2005–2007. College attendance is mensured by receipt of a 1098-T form, issued by higher education institutions to report tuition payments or scholarships. The earnings-based index of college quality is a measure of the mean earnings of all former attendees of each college in the U.S. population at age 28, as described in the text. For individuals who did not attend college, college quality is defined by mean earnings at age 28 of those who did not attend college in the U.S. population. Summary index is the standardized sum of five measures, each standardized on its own before the sum: home ownership, 401(k) retirement savings, marital status, cross-state mobility, and percent of college graduates in the individual's 2007 ZIP code of residence.

College Attendance. We begin by analyzing the impacts of class size on college attendance. Figure IIa plots the fraction of students who attend college in each year from 1999 to 2007 by class size. In this and all subsequent figures, we adjust for school-by-entry-grade effects to isolate the random variation of interest. To do so, we regress the outcome variable on school-by-entry-grade dummies and the small class indicator in each tax year. We then construct the two series shown in the figure by setting the difference between the two lines equal to the regression coefficient on the small class indicator in the corresponding year and the weighted average of the lines equal to the sample average in that year.

Figure 2.2a shows that students assigned to a small class are more likely to attend college, particularly before age 25. As the cohort ages from 19 (in 1999) to 27 (in 2007), the attendance rate of both treatment and control students declines, consistent with patterns in the broader U.S. population. Because our measure of college attendance is based on tuition payments, it includes students who attend higher education institutions both part-time and full-time. Measures of college attendance around age 20 (two years after the expected date of high school graduation) are most likely to pick up full-time attendance to two-year and four-year colleges, while college attendance in later years may be more likely to reflect part-time enrollment. This could explain why the effect of class size becomes much smaller after age 25. We therefore analyze two measures of college attendance below: college attendance at age 20 and attendance at any point before age 27.

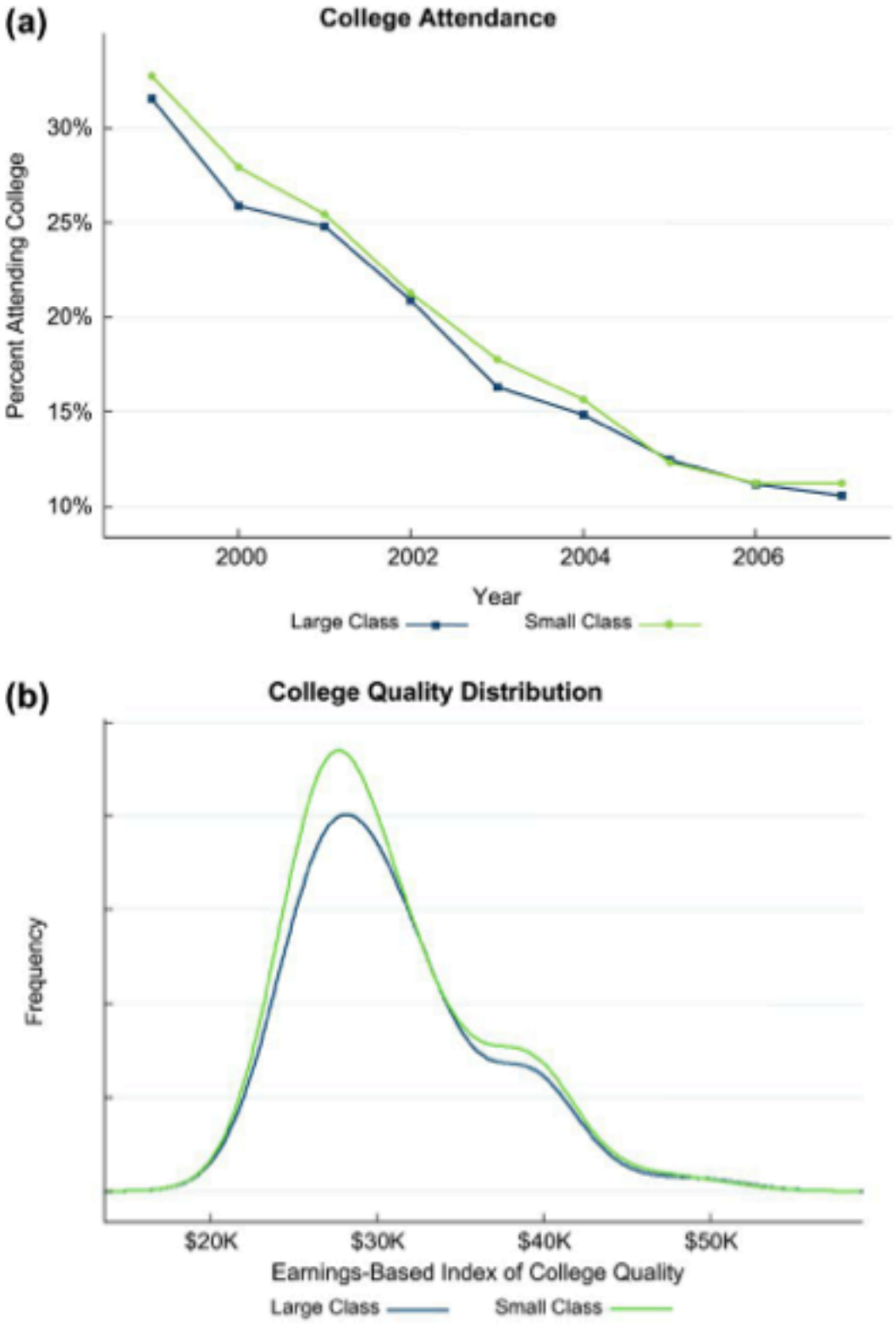

(a)
College Attendance
Percent Attending College
30%
25%
20%
15%
10%
2000
2002
2004
2006
Year
Large Class
Small Class
(b)
College Quality Distribution
Frequency
$20K
$30K
$40K
$50K
Earnings-Based Index of College Quality
Large Class
Small Class

Figure 2.2

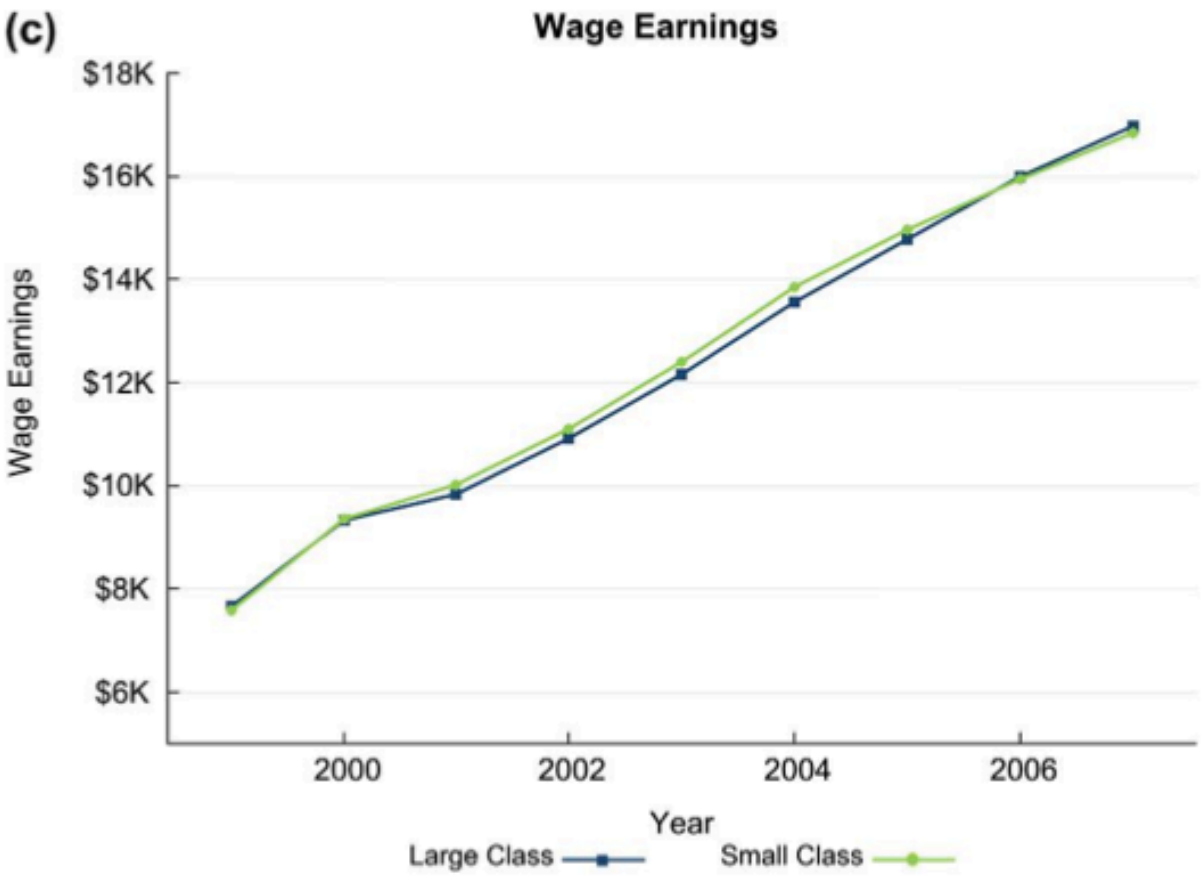


Figure 2.2 (Continued)

Effects of Class Size

Panels (a) and (c) show college attendance rates and mean wage earnings by year (from ages 19 to 27) for students randomly assigned to small and large classes. Panel (b) plots the distribution of college quality attended in 2000 using the earnings-based college quality index described in Figure Ic. Individuals who did not attend college are included in Panel (b) with college quality defined as mean earnings in the U.S. population for those who did not attend college. Kernel-smoothed densities in Panel (b) are scaled to integrate to total attendance rates for both small and large classes. All figures adjust for school-by-entry-grade effects to isolate the random variation in class size. In (a) and (c), we adjust for school-by-entry-grade effects by regressing the outcome variable on school-by-entry-grade dummies and the small class indicator in each tax year. We then construct the two series shown in the figure by requiring that the difference between the two lines equals the regression coefficient on the small class indicator in the corresponding year and that the weighted average of the lines equals the sample average in that year. In (b), we compute residual college mean earnings from a regression on school-by-entry-grade effects and plot the distribution of the residual within small and large classes, adding back the sample mean to facilitate interpretation of units. See notes to Figure I for definitions of wage earnings and college variables.

The regression estimates reported in Column 2 of Table 2.5 are consistent with the results in Figure 2.2a. Controlling for demographic characteristics, students assigned to a small class are 1.8 percentage points (6.7%) more likely to attend college in 2000. This effect is marginally significant with $p = 0.06$. Column 3 shows that students in small classes are 1.6 percentage points more likely to attend college at some point before age 27.

Next, we investigate how class size affects the quality of colleges that students attend. Using the earnings-based college quality measure described above, we plot the distribution of college quality attended in 2000 by small and large class assignment in Figure 2.2b. We compute residual college mean earnings from a regression on school-by-entry-grade effects and plot the distribution of the residuals within small and large classes, adding back the sample mean to facilitate interpretation of units. To show where the excess density in the small class group lies, the densities are scaled to integrate to the total college attendance rates for small and large classes. The excess density in the small class group lies primarily among the lower quality colleges, suggesting that the marginal students who were induced to attend college because of reduced class size enrolled in relatively low quality colleges.

Column 4 of Table 2.5 shows that students assigned to a small class attend colleges whose students have mean earnings that are $109 higher. That is, based on the cross-sectional relationship between earnings and attendance at each college, we predict that students in small classes will be earning approximately $109 more per year at age 28. This earnings increase incorporates the extensive-margin of higher college attendance rates, because students who do not attend college are assigned the mean earnings of individuals who do not attend college in our index.[59] Conditional on attending college, students in small classes attend *lower* quality colleges on average because of the selection effect shown in Figure 2.2b.[60]

Earnings. Figure 2.2c shows the analog of Figure IIa for wage earnings. Earnings rise rapidly over time because many students are in college in the early years of the sample. Individuals in small classes have slightly higher earnings than those in large classes in most years. Column 5 of

[59] Alternative earnings imputation procedures for those who do not attend college yield similar results. For example, assigning these students the mean earnings of Tennessee residents or STAR participants who do not attend college generates larger estimates.

[60] Because of the selection effect, we are unable to determine whether there was an intensive-margin improvement in quality of college attended. Quantifying the effect of reduced class size on college quality for those who were already planning to attend college would require additional assumptions such as rank preservation.

Table 2.5 shows that without controls, students who were assigned to small classes are estimated to earn $4 more per year on average between 2005 and 2007. With controls for demographic characteristics, the point estimate of the earnings impact becomes -$124 (with a standard error of $336). Though the point estimate is negative, the upper bound of the 95% confidence interval is an earnings gain of $535 (3.4%) gain per year. If we were to predict the expected earnings gain from being assigned to a small class from the cross-sectional correlation between test scores and earnings reported in column 4 of Table 2.4, we obtain an expected earnings effect of 4.8 percentiles $\times$ \$90 = \$432. This prediction lies within the 95% confidence interval for the impact of class size on earnings. In Online Appendix Table IX, we consider several alternative measures of earnings, such as total household income and an indicator for positive wage earnings. We find qualitatively similar impacts – point estimates close to zero with confidence intervals that include the predicted value from cross-sectional estimates – for all of these measures. We conclude that the class size intervention, which raises test scores by 4.8 percentiles, is unfortunately not powerful enough to detect earnings increases of a plausible magnitude as of age 27. Because class size has impacts on college attendance, earnings effects might emerge in subsequent years, especially since college graduates have much steeper earnings profiles than non college graduates.

Other Outcomes. Column 6 of Table 2.5 shows that students assigned to small classes score 4.6 percent of a standard deviation higher in the summary outcome index defined in Section III, an effect that is statistically significant with $p < 0.05$. This index combines information on savings behavior, home ownership, marriage rates, mobility rates, and residential neighborhood quality. In Online Appendix Table X, we analyze the impacts of class size on each of the five outcomes separately. We find particularly large and significant impacts on the probability of having a 401(k), which can be thought of as a proxy for having a good job. This result is consistent with the view that students in small classes may have higher permanent income that could emerge in wage earnings measures later in their lifecycles. We also find positive effects on all the other components of the summary index, though these effects are not individually significant.[61]

In Online Appendix Table XI, we document the heterogeneity of class size impacts across

[61] In Online Appendix Table X, we also analyze an alternative summary index that weights each of the five components by their impacts on wage earnings. We construct this index by regressing wage earnings on the five components in the cross-section and predicting wage earnings for each individual. We find significant impacts of class size on this predicted-earnings summary index, confirming that our results are robust to the way in which the components of the summary index are weighted.

subgroups. We replicate the analysis of class size impacts in Table 2.5 for six groups: black and white students, males and females, and lower- and higher-income students (based on free lunch status). The point estimates of the impacts of class size are positive for most of the groups and outcomes. The impacts on adult outcomes are somewhat larger for groups that exhibit larger test scores increases. For instance, black students assigned to small classes score 6.9 percentile points higher on their entry-grade test, are 5.3 percentage points more likely to ever attend college, and have an earnings increase of $250 (with a standard error of $540). There is some evidence that reductions in class size may have more positive effects for men than women and for higher income than lower income (free-lunch eligible) students. Overall, however, the STAR experiment is not powerful enough to detect heterogeneity in the impacts of class size on adult outcomes with precision.

II.D.2 Observable Teacher and Peer Effects

We estimate the impacts of observable characteristics of teachers and peers using specifications analogous to equation (10):

$$y_{icnw} = \alpha_{nw} + \beta_1 \text{SMALL}_{cnw} + \beta_2 z_{cnw} + X_{icnw}\delta + \varepsilon_{icnw} \tag{11}$$

where z_{cnw} denotes a vector of teacher or peer characteristics for student i assigned to classroom c at school n in entry grade w. Because students and teachers were randomly assigned to classrooms, β_2 can be interpreted as the effect of the relevant teacher or peer characteristics on the outcome y. Note that we control for class size in these regressions, so the variation identifying teacher and peer effects is orthogonal to that used above.

Teachers. We begin by examining the impacts of teacher experience on scores and earnings. Figure 2.3a plots KG scores vs. the numbers of years of experience that the student's KG teacher had at the time she taught his class. We exclude students who entered the experiment in grades 1 to 3 in these graphs for reasons we discuss below. We adjust for school effects by regressing the outcome and dependent variables on these fixed effects and computing residuals. The figure is a scatter plot of the residuals, with the sample means added back in to facilitate interpretation of the axes. Figure 2.3a shows that students randomly assigned to more experienced KG teachers

have higher test scores. The effect of experience on KG scores is roughly linear in the STAR experimental data, in contrast with other studies which find that the returns to experience drop sharply after the first few years.

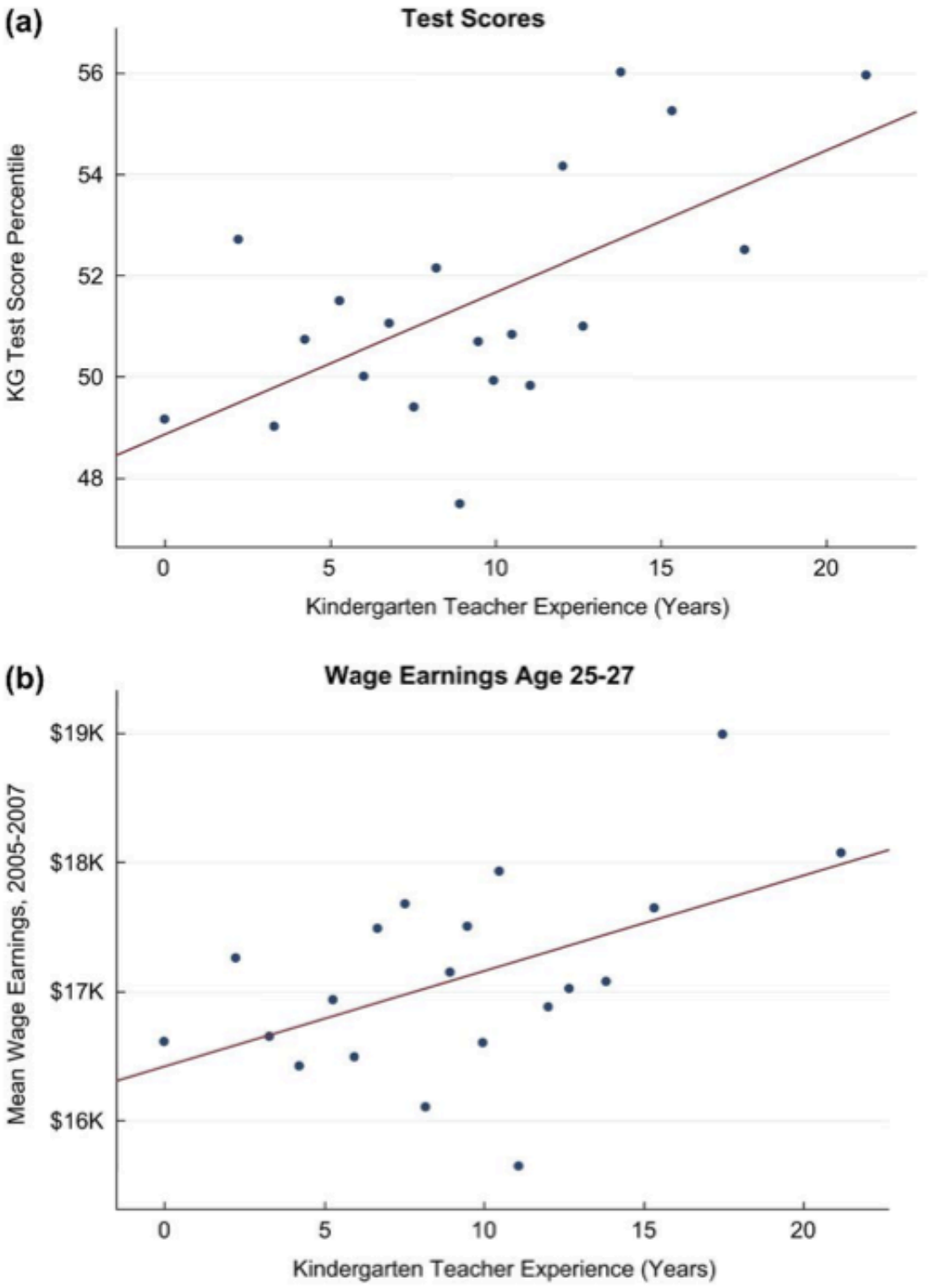

(a)
Test Scores
KG Test Score Percentile
56
54
52
50
48
0
5
10
15
20
Kindergarten Teacher Experience (Years)
(b)
Wage Earnings Age 25-27
Mean Wage Earnings, 2005-2007
$19K
$18K
$17K
$16K
0
5
10
15
20
Kindergarten Teacher Experience (Years)

Figure 2.3

Figure 2.3 (Continued)

Effects of Teacher Experience

Panel (a) plots kindergarten average test scores in math and reading (measured by within-sample percentile ranks) vs. kindergarten teacher's years of prior experience. Panel (b) plots mean wage earnings over years 2005-2007 vs. kindergarten teacher's years of prior experience. In both Panels (a) and (b), we bin teacher experience into twenty equal sized (5 percentile point) bins and plot the mean of the outcome variable within each bin. The solid line shows the best linear fit estimated on the underlying student-level data using OLS. Panel (c) plots mean wage earnings by year (from ages 19 to 27) for individuals who had a teacher with fewer than 10 or more than 10 years of experience in kindergarten. All figures adjust for school-by-entry-grade effects to isolate the random variation in teacher experience. In (a) and (b), we adjust for school-by-entry-grade effects by regressing both the dependent and independent variables on school-by-entry-grade dummies. We then plot the residuals, adding back the sample means to facilitate interpretation of units. The solid line shows the best linear fit estimated on the underlying data using OLS. In (c), we follow the same procedure used to construct Figure IIc. See notes to Figure I for definition of wage earnings.

Figure 2.3b replicates 2.3a for the earnings outcome. It shows that students who were randomly assigned to more experienced KG teachers have higher earnings at age 27. As with scores, the impact of experience on earnings in these data appear roughly linear. Figure 2.3c characterizes the time path of the earnings impact. We divide teachers in two groups – those with experience above and below 10 years (since mean years of experience is 9.3 years). We then plot mean earnings for the students in the low- and high-experience groups by year, adjusting for school fixed effects as in Figure 2.3b. From 2000 to 2004 (when students are aged 20 to 24), there is little difference in earnings between the two curves. A gap opens starting in 2005; by 2007, students who had high-experience teachers in kindergarten are earning $1,104 more on average.

Columns 1-2 of Table 2.6 quantify the impacts of teacher experience on scores and earnings, conditioning on the standard vector of student and parent demographic characteristics as well as whether the teacher has a master's degree or higher and the small class indicator. Column 1 shows that students assigned to a teacher with more than 10 years of experience score 3.2 percentile points higher on KG tests. Column 2 shows that these same students earn $1,093 more on average between ages 25 and 27 ($p < 0.05$).[62]

[62] In Online Appendix Table XII, we replicate columns 1 and 2 for small and large classes separately to evaluate whether teacher experience is more important in managing classrooms with many students. We find some evidence that teacher experience has a larger impact on earnings in large classes, but the difference in impacts is not statistically significant.

Table 2.6

OBSERVABLE TEACHER AND PEER EFFECTS

Dependent variable	(1) Test score (%)	(2) Wage earnings ($)	(3) Test score (%)	(4) Wage earnings ($)	(5) Test score (%)	(6) Wage earnings (%)	(7) Wage earnings (%)
Teacher with >10 years of experience	3.18	1093	1.61	−536.1			
	(1.26)	(545.5)	(1.21)	(619.3)			
Teacher has post-BA deg.	−0.848	−261.1	0.95	−359.4			
	(1.15)	(449.4)	(0.90)	(500.1)			
Fraction black classmates					−6.97	−1,757	
					(9.92)	(2692)	
Fraction female classmates					9.74	-67.53	
					(4.26)	(1539)	
Fraction free-lunch classmates					−7.53	−284.6	
					(4.40)	(1731)	
Classmates' mean age					−3.24	-25.78	
					(3.33)	(1359)	
Classmates' mean predicted score							−23.06
							(94.07)
Small class	5.19	−8.158	3.77	−284.2	4.63	-132.2	−119.2
	(1.19)	(448.4)	(1.17)	(536.4)	(0.99)	(342.3)	(330.9)
Entry grade	KG	KG	Grade ≥1	Grade ≥1	All	All	All
Observations	5,601	6,005	4,270	4,909	9,939	10,992	10,992

Notes. Each column reports coefficients from an OLS regression, with standard errors clustered by school in parentheses. All specifications control for school-by-entry-grade fixed effects, an indicator for initial assignment to a small class, and the vector of demographic characteristics used first in TableIV: a quartic in parent's household income interacted with an indicator for whether the filing parent is ever married between 1996 and 2008, mother's age at child's birth, indicators for parent's 401(k) savings and home ownership, and student's race, gender, free-lunch status, and age at kindergarten. Columns (1)–(2) include only students who entered a STAR school in kindergarten. Columns (3)–(4) include only students who enter STAR after kindergarten. Columns (5)–(7) pool all students, regardless of their entry grade. Test score is the average math and reading test score at the end of the year in which the student enters a STAR school (measured in percentiles). Wage earnings is the individual's mean wage earnings over years 2005–2007 (including 0s for people with no wage earnings). Teacher experience is the number of years the teacher taught at any school before the student's year of entry into a STAR school. Classmates' characteristics are defined based on the classroom that the student enters in the first year he is in a STAR school and omit the own student. Classmates' mean predicted score is constructed by regressing test scores on school-by-entry-grade fixed effects and the demographic characteristics listed above and then taking the mean of the predicted scores. Variables labeled "fraction" are in units between 0 and 1. Free-lunch status is a proxy for having low household income.

Columns 3-4 show that teacher experience has a much reduced effect for children entering the experiment in grades 1 to 3 on both test scores and earnings. The effect of teacher experience on test scores is no longer statistically significant in grades 1-3. Consistent with this result, teacher experience in grades 1-3 also does not have a statistically significant effect on wage earnings. Unfortunately, the STAR dataset includes very few teacher characteristics, so we are unable to provide definitive evidence on why the effect of teacher experience varies across grades.

The impact of kindergarten teacher experience on earnings must be interpreted very carefully. Our results show that placing a child in a kindergarten class taught by a more experienced teacher yields improved outcomes. This finding does *not* imply that increasing a given teacher's experience will improve student outcomes. The reason is that while teachers were randomly assigned to classrooms, experience was not randomly assigned to teachers. The difference in earnings of students with experienced teachers could be due to the intrinsic characteristics of experienced teachers rather than experience of teachers per se. For instance, teachers with more experience have selected to stay in the profession and may be more passionate or more skilled at teaching. Alternatively, teachers from older cohorts may have been more skilled (Corcoran, Evans, and Schwab 2004, Hoxby and Leigh 2004, Bacolod 2007). These factors may explain the difference between the effect of teacher experience in Kindergarten and later grades. For instance, the selection of teachers may vary across grades or cohort effects may differ for Kindergarten teachers.

The linear relationship between kindergarten teacher experience and scores in the STAR data stands in contrast to earlier studies that track teachers over time in a panel and find that teacher performance improves with the first few years of experience and then plateaus. This further suggests that other factors correlated with experience may drive the observed impacts on scores and earnings. We therefore conclude that early childhood teaching has a causal impact on long term outcomes but we cannot isolate the characteristics of teachers responsible for this effect.

The few other observable teacher characteristics in the STAR data (degrees, race, and progress on a career ladder) have no significant impact on scores or earnings. For instance, columns 1-4 of Table 2.6 show that the effect of teachers' degrees on scores and earnings is statistically insignificant. The finding that experience is the only observable measure that predicts teacher quality matches earlier studies of teacher effects (Hanushek 2010, Rockoff and Staiger 2010).[63]

[63] Dee (2004) shows that being assigned to a teacher of the same race raises test scores. We find a positive but

Peers. Better classmates could create an environment more conducive to learning, leading to improvements in adult outcomes. To test for such peer effects, we follow the standard approach in the recent literature by using linear-in-means regressions specifications. We include students who enter in all grades and measure peer characteristics in their first, randomly assigned classroom, and condition on school-by-entry-grade effects. We proxy for peer abilities (z) in equation (11) with the following exogenous peer characteristics: fraction black, fraction female, fraction eligible for free or reduced-price lunch (a proxy for low income), and mean age. Replicating previous studies, we show in column 5 of Table 2.6 that the fraction of female and low-income peers significantly predict test scores. Column 6 replicates column 5 with earnings as the dependent variable. The estimates on all four peer characteristics are very imprecise. For instance, the estimated effect of increasing the fraction of low-income peers by 10 percentage points is an earnings loss of $28, but with a standard error of $173. In an attempt to obtain more power, we construct a single index of peer abilities by first regressing scores on the full set of parent and student demographic characteristics described above and then predicting peers' scores using this regression. However, as column 7 shows, even the predicted peer score measure does not yield a precise estimate of peer effects on earnings; the 95% confidence interval for a 1 percentile point improvement in peers' predicted test scores ranges from -$207 to $160.[64]

The STAR experiment lacks the power to measure the effects of observable peer characteristics on earnings precisely because the experimental design randomized students across classrooms. As a result, it does not generate significant variation in mean peer abilities across classes. The standard deviation of mean predicted peer test scores (removing variation across schools and waves) is less than two percentile points. This small degree of variation in peer abilities is adequate to identify some contemporaneous effects on test scores but proves to be insufficient to identify effects on outcomes twenty years later, which are subject to much higher levels of idiosyncratic noise.

II.E Impacts of Unobservable Classroom Characteristics

Many unobserved aspects of teachers and peers could impact student achievement and adult outcomes. For instance, some teachers may generate greater enthusiasm among students or some

statistically insignificant impact of having a teacher of the same race on earnings.

[64] We find positive but insignificant impacts of teacher and peer characteristics on the other outcomes above, consistent with a general lack of power in observable characteristics (not reported).

peers might be particularly disruptive. To test whether such unobservable aspects of class quality have long-term impacts, we estimate the parameters of a correlated random effects model. In particular, we test for "class effects" on scores and earnings by exploiting random assignment to classrooms. These class effects include the effects of teachers, peers, and any class-level shocks. We formalize our estimation strategy using a simple empirical model.

II.E.1 A Model of Class Effects

For simplicity, we analyze a model in which all students enter in the same grade and suppress the entry grade index (w); we discuss below how our estimator can be applied to the case with multiple entry grades. We first consider a case without peer effects and then show how peer effects affect our analysis below.

Consider the following model of test scores (s_{icn}) at the end of the class and earnings or other adult outcomes (y_{icn}) for student i in class c at school n:

$$s_{icn} = d_n + \sum_k \mu_k^S Z_{cn}^k + a_{icn} \tag{12}$$

$$y_{icn} = \delta_n + \sum_k \mu_k^Y Z_{cn}^k + \rho a_{icn} + \nu_{icn}, \tag{13}$$

where the error term a_{icn} can be interpreted as intrinsic academic ability. The error term ν_{icn} represents the component of intrinsic earnings ability that is uncorrelated with academic ability. The parameter ρ controls the correlation between intrinsic academic and earnings ability. The school fixed effects d_n and δ_n capture school-level differences in achievement on tests and earnings outcomes, e.g. due to variation in socioeconomic characteristics across school areas. $Z_{cn} = (Z_{cn}^1, .., Z_{cn}^K)$ denotes a vector of classroom characteristics such as class size, teacher experience, or other teacher attributes. The coefficients μ_k^S and μ_k^Y are the effects of class characteristic k on test scores and earnings respectively. Note that the ratios of μ_k^Y/μ_k^S may vary across characteristics. For example, teaching to the test could improve test scores but not earnings, while an inspiring teacher who does not teach to the test might raise earnings without improving test scores.

Denote by $z_{cn} = \sum_k \mu_k^S Z_{cn}^k$ the total impact of the bundle of class characteristics offered in classroom c on scores. The total impact of classrooms on earnings can be decomposed as $\sum_k \mu_k^Y Z_{cn}^k = \beta z_{cn} + z_{cn}^Y$, where z_{cn}^Y is by construction orthogonal to z_{cn}. Hence, we can rewrite

equations (12) and (13) as

$$s_{icn} = d_n + z_{cn} + a_{icn} \tag{14}$$

$$y_{icn} = \delta_n + \beta z_{cn} + z^Y_{cn} + \rho a_{icn} + \nu_{icn}. \tag{15}$$

In this correlated random effects model, z_{cn} represents the component of classrooms that affects test scores (and earnings if $\beta > 0$), while z^Y_{cn} represents the component of classrooms that affects only earnings without affecting test scores. Class effects on earnings are determined by both β and $var(z^Y_{cn})$. The parameter β measures the correlation of class effects on scores and class effects on earnings. Importantly, β only measures the impact of the bundle of classroom-level characteristics that varied in the STAR experiment rather than the impact of any single characteristic. Because β is not a structural parameter, not all educational interventions that improve test scores will have the same effect on earnings.[65] Moreover, we could find $\beta > 0$ even if no single characteristic affects both test scores and earnings.[66]

Because of random assignment to classrooms, students' intrinsic abilities a_{icn} and ν_{icn} are orthogonal to z_{cn} and z^Y_{cn}. Exploiting this orthogonality condition, one can estimate equations (12) and (13) directly using OLS for characteristics that are directly observable, as we did using equations (10) and (11) to analyze the impacts of class size and observable teacher and peer attributes. To analyze unobservable attributes of classrooms, we use two techniques: an analysis of variance to test for class effects on earnings ($\beta var(z_{cn}) + var(z^Y_{cn}) > 0$) and a regression-based method to test for covariance of class effects on scores and earnings ($\beta > 0$).

Analysis of Variance: class effects on scores and earnings. We decompose the variation in y_{icn} into individual and class-level components and test for the significance of class-level variation using an ANOVA. Intuitively, the ANOVA tests whether the outcome y varies across classes by more than what would be predicted by random variation in students across classrooms. We measure the magnitude of the class effects on earnings using a random class effects specification for equation (15) to estimate the standard deviation of class effects under the assumption that they are normally

[65] As an extreme example, teachers who help students raise test scores by cheating may have zero impact on earnings. The β estimated below applies to the set of classroom characteristics that affected test scores in the STAR experiment.

[66] Suppose teaching to the test affects only test scores while teaching discipline affects only earnings. If the decisions of teachers to teach to the test and teach discipline are correlated, then we would still obtain $\beta > 0$ in (15).

distributed.

Although the ANOVA is useful for estimating the magnitude of class effects on earnings, it has two limitations. First, it does not tell us whether class effects on scores are correlated with class effects on earnings (i.e., whether $\beta > 0$). Hence, it does not answer a key question: do classroom environments that raise test scores also improve adult outcomes? This is an important question because the impacts of most educational policies can be measured only by test scores in the short run. Second, in the STAR data, roughly half the students enter in grades 1-3 and are randomly assigned to classrooms at that point. Because only a small number of students enter each school in each of these later grades, we do not have the power to detect class effects in later grades and therefore do not include these students in the ANOVA.

Covariance between class effects on scores and earnings. Motivated by these limitations, our second strategy measures the covariance between class effects on scores and class effects on earnings (β). As the class effect on scores z_{cn} is unobserved, we proxy for it using end-of-class peer test scores. Let s_{cn} denote the mean test score in class c (in school n) and s_n denote the mean test score in school n. Let I denote the number of students per class, C the number of classes per school, and N the number of schools.[67] The mean test score in class c is

$$s_{cn} = \frac{1}{I}\sum_{i=1}^{I} s_{icn} = d_n + z_{cn} + \frac{1}{I}\sum_{i=1}^{I} a_{icn}$$

To simplify notation, assume that the mean value of z_{cn} across classes within a school is 0 ($z_n = 0$). Then the difference between mean test scores in class c and mean scores in the school is

$$\Delta s_{cn} = s_{cn} - s_n = z_{cn} + \left[\frac{1}{I}\sum_{j=1}^{I} a_{jcn} - \frac{1}{IC}\sum_{c=1}^{C}\sum_{j=1}^{I} a_{jcn}\right]. \tag{16}$$

Equation (16) shows that Δs_{cn} is a (noisy) observable measure of class quality z_{cn}. The noise arises from variation in student abilities across classes. As the number of students grows large ($I \to \infty$), Δs_{cn} converges to the true underlying class quality z_{cn} if all students are randomly assigned to classrooms.

[67] We assume that I and C do not vary across classes and schools for presentational simplicity. Our empirical analysis accounts for variation in I and C across classrooms and schools, and the analytical results below are unaffected by such variation.

Equation (16) motivates substituting Δs_{cn} for z_{cn} in equation (15) and estimating a regression of the form:

$$y_{icn} = \alpha_n + b^M \Delta s_{cn} + \varepsilon_{icn}. \tag{17}$$

The OLS estimate $\widehat{b^M}$ is a consistent estimate of β as the number of students $I \to \infty$, but it is upward-biased with finite class size because a high ability student raises the average class score and also has high earnings himself. Because of this own-observation problem, $plim_{\ N \longrightarrow \infty} \hat{b}^M > 0$ even when $\beta = 0$ (see Online Appendix B). An intuitive solution to eliminate the upward bias due to the own-observation problem is to omit the own score s_{icn} from the measure of class quality for individual i. Hence, we proxy for class quality using a leave-out mean (or jackknife) peer score measure

$$\Delta s_{cn}^{-i} = s_{cn}^{-i} - s_{n}^{-i}, \tag{18}$$

where

$$s_{cn}^{-i} = \frac{1}{I-1} \sum_{j=1, j \neq i}^{I} s_{jcn}$$

is classmates' mean test scores and

$$s_{n}^{-i} = \frac{1}{IC-1} \sum_{k=1}^{C} \sum_{j=1, j \neq i}^{I} s_{jkn}$$

is schoolmates' mean scores. Intuitively, the measure Δs_{cn}^{-i} answers the question: "How good are your classmates' scores compared with those of classmates you could have had in your school?" Replacing Δs_{cn} by Δs_{cn}^{-i}, we estimate regressions of the following form:

$$y_{icn} = \alpha_n + b^{LM} \Delta s_{cn}^{-i} + \varepsilon_{icn}. \tag{19}$$

We show in Online Appendix B that the coefficient on class quality converges to a positive value as the number of schools N grows large if and only if class quality has an impact on adult outcomes: $plim_{\ N \longrightarrow \infty} \hat{b}^{LM} > 0$ iff $\beta > 0$.[68] However, b^{LM} is biased toward zero relative to β because Δs_{cn}^{-i} is a noisy measure of class quality. In Online Appendix B, we use the sample variance of test scores

[68] We use the difference between peer scores in the class and the school (rather than simply using classmates' scores) to address the finite-sample bias in small peer groups identified by Guryan, Kroft, and Notowidigdo (2009).

to estimate the degree of this attenuation bias at 23%.

Our preceding analysis ignores variation in class quality due to peer effects. With peer effects, a high ability student may raise his peers' scores, violating the assumption made above that $z_{cn} \perp a_{icn}$. Such peer effects bias b^{LM} upward (generating $plim\ _{N\longrightarrow\infty}\ \hat{b}^{LM} > \beta$) because of the reflection problem (Manski 1993). Even if there is no effect of class quality on earnings, that student's higher earnings (due solely to her own ability) will generate a positive correlation between peer scores and own earnings. While we cannot purge our leave-out-mean estimator of this bias, we show below that we can tightly bound the degree of reflection bias in a linear-in-means model. The reflection bias turns out to be relatively small in our application because it is of order $\frac{1}{I}$ and classes have 20 students on average.

We refer to peer-score measure Δs_{cn}^{-i} as "class quality" and the coefficient b^{LM} as the effect of class quality on earnings (or other outcomes). Although we regress outcomes on peer scores in equation (19), the coefficient b^{LM} should not be interpreted as an estimate of peer effects. Because class quality Δs_{cn}^{-i} is defined based on *end-of-class* peer scores, it captures teacher quality, peer quality, and any other class-level shocks that may have affected students systematically. End-of-class peer scores are a single index that captures all classroom characteristics that affect test scores. Equation (19) simply provides a regression-based method of estimating the correlation between random classroom effects on scores and earnings.

We include students who enter STAR in later grades when estimating equation (19) by defining Δs_{cn}^{-i} as the difference between mean end-of-year test scores for classmates and schoolmates in the student's grade in the year she entered a STAR school. To maximize precision, we include all peers (including those who had entered in earlier grades) when defining Δs_{cn}^{-i} for new entrants. Importantly, Δs_{cn}^{-i} varies randomly within schools for new entrants – who are randomly assigned to their first classroom – as it does for kindergarten entrants.[69] With this definition of Δs_{cn}^{-i}, b^{LM} measures the extent to which class quality in the initial class of entry (weighted by the entry rates across the four grades) affects outcomes.

An alternative approach to measuring the covariance between class effects on scores and earnings is to use an instrumental variables strategy, regressing earnings on test scores and instrumenting

[69] For entrants in grades 1-3, there can be additional noise in the class quality measure because students who had entered in earlier grades were not in general re-randomized across classrooms. Because such noise is orthogonal to entering student ability, it generates only additional attenuation bias.

for scores with classroom fixed effects. Because the fitted values from the first stage regression are just mean test scores by classroom, the coefficient obtained from this TSLS regression coincides with b^M when we run equation (17). The TSLS estimate of β is upward biased because the own observation is included in both mean scores and mean earnings, which is the well known weak instruments problem. The weak instruments literature has developed various techniques to deal with this bias, including (a) jackknife IV (Angrist, Imbens, and Krueger 1999), which solves the problem by omitting the own observation when forming the instrument; (b) split-sample IV (Angrist and Krueger 1995), which randomly splits classes into two and only uses mean scores in the other half of the class as an instrument; and (c) limited information maximum likelihood (LIML), which collapses the parameter space and uses maximum likelihood to obtain a consistent estimate of β. The estimator for b^{LM} in equation (19) is essentially the reduced-form of the first technique, the jackknife IV regression. We present estimates using the instrumental variable strategies in Online Appendix Table XIII to evaluate the robustness of our results.

II.E.2 Analysis of Variance

We implement the analysis of variance using regression specifications of the following form for students who enter the experiment in kindergarten:

$$y_{icn} = \alpha_n + \gamma_{cn} + X_{icn}\delta + \varepsilon_{icn} \tag{20}$$

where y_{icn} is an outcome for student i who enters class c in school n in kindergarten and γ_{cn} is the class effect on the outcome, and X_{icn} a vector of pre-determined individual background characteristics.[70]

We first estimate equation (20) using a fixed-effects specification for the class effects γ_{cn}. Under the null hypothesis of no class effects, the class dummies should not be significant because of random assignment of students to classrooms. We test this null hypothesis using an F test for whether $\gamma_{cn} = 0$ for all c, n. To quantify the magnitude of the class effects, we compute the variance of γ_{cn} by estimating equation (20) using a random-effects specification. In particular, we assume that $\gamma_{cn} \sim N(0, \sigma_c^2)$ and estimate the standard deviation of class effects σ_c.

[70] We omit γ_{cn} for one class in each school to avoid collinearity with the school effects α_n.

Table 2.7 reports p values from F tests and estimates of σ_c for test scores and earnings. Consistent with Nye, Konstantopoulos, and Hedges (2004) – who use an ANOVA to test for class effects on scores in the STAR data – we find highly significant class effects on KG test scores. Column 1 rejects the null hypothesis of no class effects on KG scores with $p < 0.001$. The estimated standard deviation of class effects on test scores is $\sigma_c = 8.77$, implying that a one standard deviation improvement in class quality raises student test scores by 8.77 percentiles (0.32 standard deviations). Note that this measure represents the impact of improving class quality by one SD of the *within-school* distribution because the regression specification includes school fixed effects.

Table 2.7

KINDERGARTEN CLASS EFFECTS: ANALYSIS OF VARIANCE

Dependent variable	(1) Grade K scores	(2) Grade 8 scores	(3)	(4) Wage earnings	(5)	(6)
p-value of *F* test on KG class fixed effects	0.000	0.419	0.047	0.026	0.020	0.040
p-value from permutation test	0.000	0.355	0.054	0.029	0.023	0.055
SD of class effects (RE estimate)	8.77%	0.000%	$1,497	$1,520	$1,703	$1,454
Demographic controls	x	x		x	x	x
Large classes only					x	
Observable class chars.						x
Observations	5,621	4,448	6,025	6,025	4,208	5,983

Notes. Each column reports estimates from an OLS regression of the dependent variable on school and class fixed effects, omitting one class fixed effect per school. The *p*-value in the first row is for an *F* test of the joint significance of the class fixed effects. The second row reports the *p*-value from a permutation test, calculated as follows: we randomly permute students between classes within each school, calculate the *F*-statistic on the class dummies, repeat the previous two steps 1,000 times, and locate the true *F*-statistic in this distribution. The third row reports the estimated standard deviation of class effects from a model with random class effects and school fixed effects. Grade 8 scores are available for students who remained in Tennessee public schools and took the eighth-grade standardized test any time between 1990 and 1997. Both KG and eighth-grade scores are coded using within-sample percentile ranks. Wage earnings is the individual's mean wage earnings over years 2005–2007 (including 0s for people with no wage earnings). All specifications are estimated on the subsample of students who entered a STAR school in kindergarten. All specifications except (3) control for the vector of demographic characteristics used in Table IV: a quartic in parent's household income interacted with an indicator for whether the filing parent is ever married between 1996 and 2008, mother's age at child's birth, indicators for parent's 401(k) savings and home ownership, and student's race, gender, free-lunch status, and age at kindergarten. Column (5) limits the sample to large classes only; this column identifies pure KG class effects because students who were in large classes were rerandomized into different classes after KG. Column (6) replicates column (4), adding controls for the following observable classroom characteristics: indicators for small class, above-median teacher experience, black teacher, and teacher with degree higher than a BA, and classmates' mean predicted score. Classmates' mean predicted score is constructed by regressing test scores on school-by-entry-grade fixed effects and the vector of demographic characteristics listed above and then taking the mean of the predicted scores.

Column 2 of Table 2.7 replicates the analysis in column 1 with 8th grade test scores as the outcome. We find no evidence that kindergarten classroom assignment has any lasting impact on achievement in 8th grade as measured by standardized test scores ($p = 0.42$). As a result, the estimated standard deviation of class effects on 8th grade scores is $\sigma_c = 0.00$. This evidence suggests that KG class effects fade out by grade 8, a finding that we revisit and explore in detail in Section 2.6.

Columns 3-6 of Table 2.7 implement the ANOVA for earnings (averaged over ages 25-27). Column 3 implements the analysis without any controls besides school fixed effects. Column 4 introduces the full vector of parental and student demographic characteristics. Both specifications show statistically significant class effects on earnings ($p < 0.05$). Recall that the same specification revealed no significant differences in *predicted* earnings (based on pre-determined variables) across classrooms ($p = 0.92$, as shown in column 6 of Table 2.2). Hence, the clustering in actual earnings by classroom is the consequence of treatments or common shocks experienced by students after random assignment to a KG classroom. The standard deviation of KG class effects on earnings in column 4 (with controls) is $\sigma_c = \$1,520$. Assigning students to a classroom that is one standard deviation better than average in kindergarten generates an increase in earnings at ages 25-27 of \$1,520 (9.6%) per year for each student. While the mean impact of assignment to a better classroom is large, kindergarten class assignment explains a small share of the variance in earnings. The intra-class correlation coefficient in earnings implied by the estimate in Column 4 of Table 2.7 is only $(1,520/15,558)^2 = 0.01$.[71]

Column 5 of Table 2.7 restricts the sample to students assigned to large classes, to test for class effects purely within large classrooms. This specification is of interest for two reasons. First, it isolates variation in class quality orthogonal to class size. Second, students in large classes were randomly reassigned to classrooms in first grade. Hence, column 5 specifically identifies clustering by kindergarten classrooms rather than a string of teachers and peers experienced over several years

[71] The clustering of earnings detected by the ANOVA may appear to contradict that fact that clustering standard errors by classroom or school has little impact on the standard errors in the regression specification in, for example, equation (10) (see Online Appendix Table VII). The intra-class correlation in earnings of 0.01 implies a Moulton correction factor of 1.09 for clustering at the classroom level with a mean class size of 20.3 students (Angrist and Pischke 2009, equation 8.2.5). The Moulton adjustment of 9% assumes that errors are equi-correlated across students within a class. Following standard practice, we report clustered standard errors that do not impose this equi-correlation assumption. Clustered standard errors can be smaller than un-clustered estimates when the intra-class correlation coefficient is small. We thank Gary Chamberlain for helpful comments on these issues.

by a group of children who all started in the same KG class. Class quality continues to have a significant impact on earnings within large classes, showing that components of kindergarten class quality beyond size matter for earnings.

Column 6 expands upon this approach by controlling for all observable classroom characteristics: indicators for small class, teacher experience above 10 years, teacher race, teacher with degree higher than a BA, and classmates' mean predicted score, constructed as in column 6 of Table 2.6. The estimated σ_c falls by only \$66 relative to the specification in column 4, implying that most of the class effects are driven by features of the classroom that we cannot observe in our data.

The F tests in Table 2.7 rely on parametric assumptions to test the null of no class effects. As a robustness check, we run permutation tests in which we randomly permute students between classes within each school. For each random permutation, we calculate the F statistic on the class dummies. Using the empirical distribution of F statistics from 1,000 within-school permutations of students, we calculate a non-parametric p value based on where the true F statistic (from row 1) falls in the empirical distribution. Reassuringly, these non-parametric p values are quite similar to those produced from the parametric F test, as shown in the second row of Table 2.7.

II.E.3 Covariance between Class Effects on Scores and Earnings

Having established class effects on both scores and earnings, we estimate the covariance of these class effects using regression specifications of the form

$$y_{icnw} = \alpha_{nw} + \beta \Delta s_{cnw}^{-i} + X_{icnw}\delta + \varepsilon_{icnw}, \tag{21}$$

where y_{icnw} represents an outcome for student i who enters class c in school n in entry grade (wave) w. The regressor of interest Δs_{cnw}^{-i} is our leave-out mean measure of peer test scores for student i at the end of entry grade w, as defined in equation (18).[72] In the baseline specifications, we include students in all entry grades to analyze how the quality of the student's randomly assigned first class affects long-term outcomes. We then test for differences in the impacts of class quality across grades K-3 by estimating equation (21) for separate entry grades. As above, we cluster standard errors at the school level to adjust for the fact that outcomes are correlated across students within

[72] Sacerdote (2001) employs analagous regression specifications to detect clustering in randomly assigned roommates' ex-post test scores.

classrooms and possibly within schools.

We begin by characterizing the impact of class quality on test scores. Figure 2.4a plots each student's end-of-grade test scores vs. his entry-grade class quality, as measured by his classmates' test scores minus his schoolmates' test scores. The graph adjusts for school-by-entry-grade effects to isolate the random variation in class quality using the technique in Figure IIIa; it does not adjust for parent and student controls. Figure 2.4a shows that children randomly assigned to higher quality classes upon entry – i.e., classes where their peers score higher on the end of year test – have higher test scores at the end of the year. A one percentile increase in entry-year class quality is estimated to raise own test scores by 0.68 percentiles, confirming that test scores are highly correlated across students within a classroom. Figure 2.4b replicates Figure 2.4a, changing the dependent variable to 8th grade test score. Consistent with the earlier ANOVA results, the impact fades out by grade 8. A one percentile increase in the quality of the student's entry-year classroom raises 8th grade test scores by only 0.08 percentiles. Figure 2.4c uses the same design to evaluate the effects of class quality on adult wage earnings. Students assigned to a one percentile higher quality class have \$56.6 (0.4%) higher earnings on average over ages 25-27.

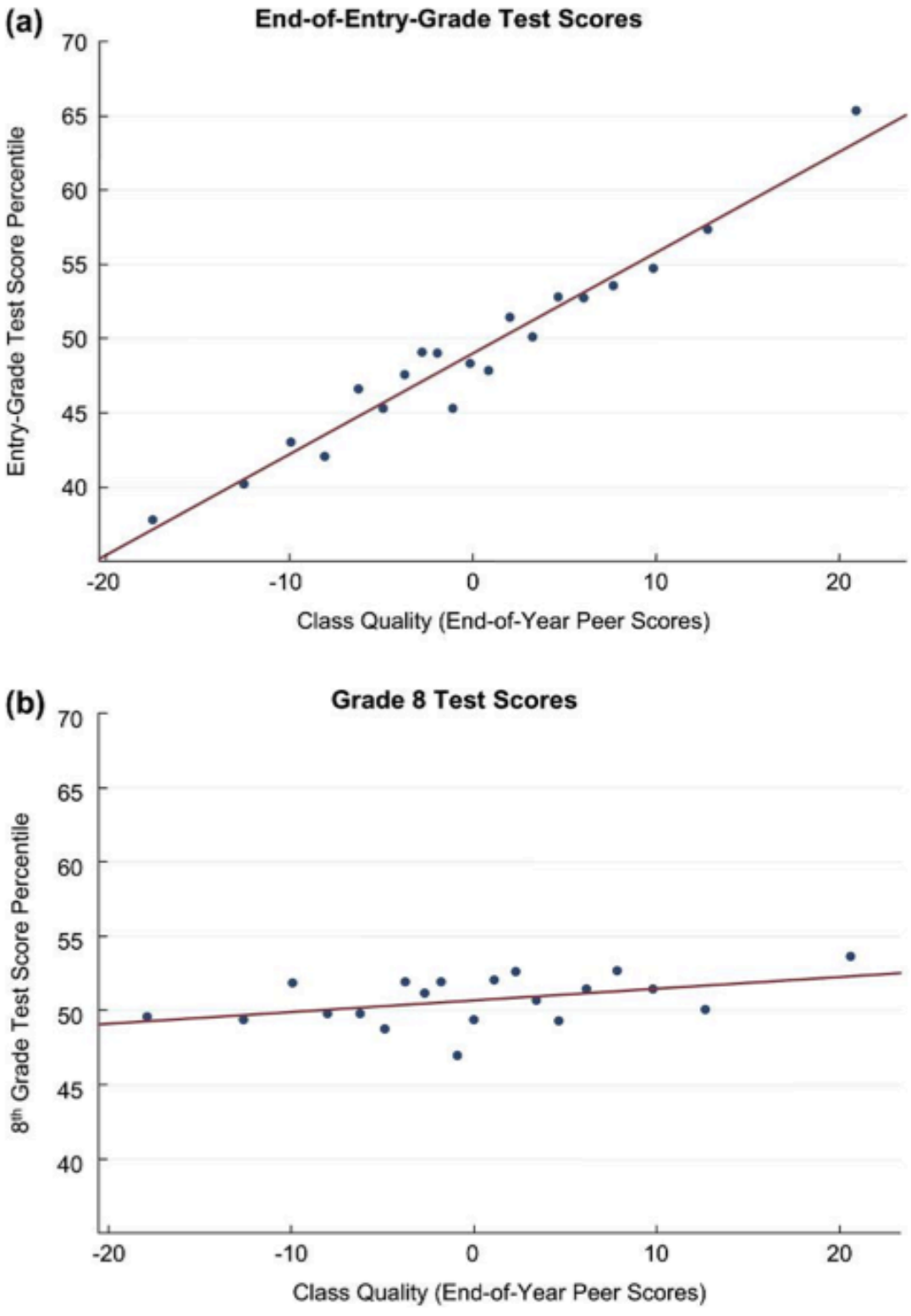

(a)
End-of-Entry-Grade Test Scores
Entry-Grade Test Score Percentile
70
65
60
55
50
45
40
-20
-10
0
10
20
Class Quality (End-of-Year Peer Scores)
(b)
Grade 8 Test Scores
8th Grade Test Score Percentile
70
65
60
55
50
45
40
-20
-10
0
10
20
Class Quality (End-of-Year Peer Scores)

Figure 2.4

Figure 2.4 (Continued)

Effects of Class Quality

The x axis in all panels is class quality, defined as the difference between the mean end-of-entry-grade test scores of a student's classmates and (grade-specific) schoolmates. Class quality is defined based on the first, randomly assigned STAR classroom (i.e., KG classroom for KG entrants, first grade classroom for 1st grade entrants, etc.). In all panels, we bin class quality into twenty equal sized (5 percentile point) bins and plot the mean of the outcome variable within each bin. The solid line shows the best linear fit estimated on the underlying student-level data using OLS. The dependent variable in Panel (a) is the student's own test score at the end of the grade in which he entered STAR. The coefficient of end-of-entry-grade test scores on class quality is 0.68 (s.e. = 0.03), implying that a 1 percentile improvement in class quality is associated with a 0.68 percentile improvement in test scores. The dependent variable in Panel (b) is a student's test score at the end of 8th grade. The coefficient of 8th grade test scores on class quality is 0.08 (s.e. = 0.03). The dependent variable in Panel (c) is a student's mean wage earnings over years 2005–2007. The coefficient of wage earnings on class quality is \$57.6 (s.e. = \$16.2), implying that a 1 percentile improvement in class quality leads to a \$57.6 increase in a student's annual earnings. All panels adjust for school-by-entry-grade effects to isolate the random variation in class quality using the technique in Figure IIIa. See notes to Figure I for definition of wage earnings.

We verify that our method of measuring class quality does not generate a mechanical correlation between peers scores and own outcomes using permutation tests. We randomly permute students across classrooms within schools and replicate equation (21). We use the t statistics on β from the random permutations to form an empirical cdf of t statistics under the null hypothesis of no class effects. We find that fewer than 0.001% of the t statistics from the random permutations are larger than the actual t statistic on kindergarten test score in Figure 2.4a of 22.7. For the earnings outcome, fewer than 0.1% of the t statistics from the random permutations are larger than the actual t statistic of 3.55. These non-parametric permutation tests confirm that the p values obtained using parametric t-tests are accurate in our application.

As noted above, part of the relationship between earnings and peers' test scores may be driven by reflection bias: high ability students raise their peers' scores and themselves have high earnings. This could generate a correlation between peer scores and own earnings even if class quality has no causal impact on earnings. However, the fact that end-of-kindergarten peer scores are not highly correlated with 8th grade test scores (Figure 2.4b) places a tight upper bound on the degree of this bias. In the presence of reflection bias, a high ability student (who raises her classroom peers' scores in the year she enters) should also score highly on 8th grade tests, creating a spurious correlation between first-classroom peer scores and own 8th grade scores. Therefore, if first-classroom peer scores have zero correlation with 8th grade scores, there cannot be any reflection bias. In Online Appendix B, we formalize this argument by deriving a bound on the degree of reflection bias in a linear-in-means model as a function of the empirical estimates in Table 2.8 and the cross-sectional correlations between test scores and earnings. If class quality has no causal impact on earnings ($\beta = 0$), the upper bound on the regression coefficient of earnings on class quality is \$9, less than 20% of our empirical estimate of \$56.6. Although this quantitative bound relies on the parametric assumptions of a linear-in-means model, it captures a more general intuition: the rapid fade out of class quality effects on test scores rules out significant reflection bias in impacts of peer scores on later adult outcomes. Recall that the class quality estimates also suffer from a downward attenuation bias of 23%, the same magnitude as the upper bound on the reflection bias. We therefore proceed by using end-of-year peer scores as a simple proxy for class quality.

Figure 2.5a characterizes the time path of the impact of class quality on earnings, dividing classrooms in two groups – those with class quality above and below the median. The time

pattern of the total class quality impact is similar to the impact of teacher experience shown in Figure 2.3c. Prior to 2004, there is little difference in earnings between the two curves, but the gap noticeably widens beginning in 2003. By 2007, students who were assigned to classes of above-median quality are earning $875 (5.5%) more on average. Figure 2.5b shows the time path of the impacts on college attendance. Students in higher quality classes are more likely to be attending college in their early 20's, consistent with their higher earnings and steeper earnings trajectories in later years.

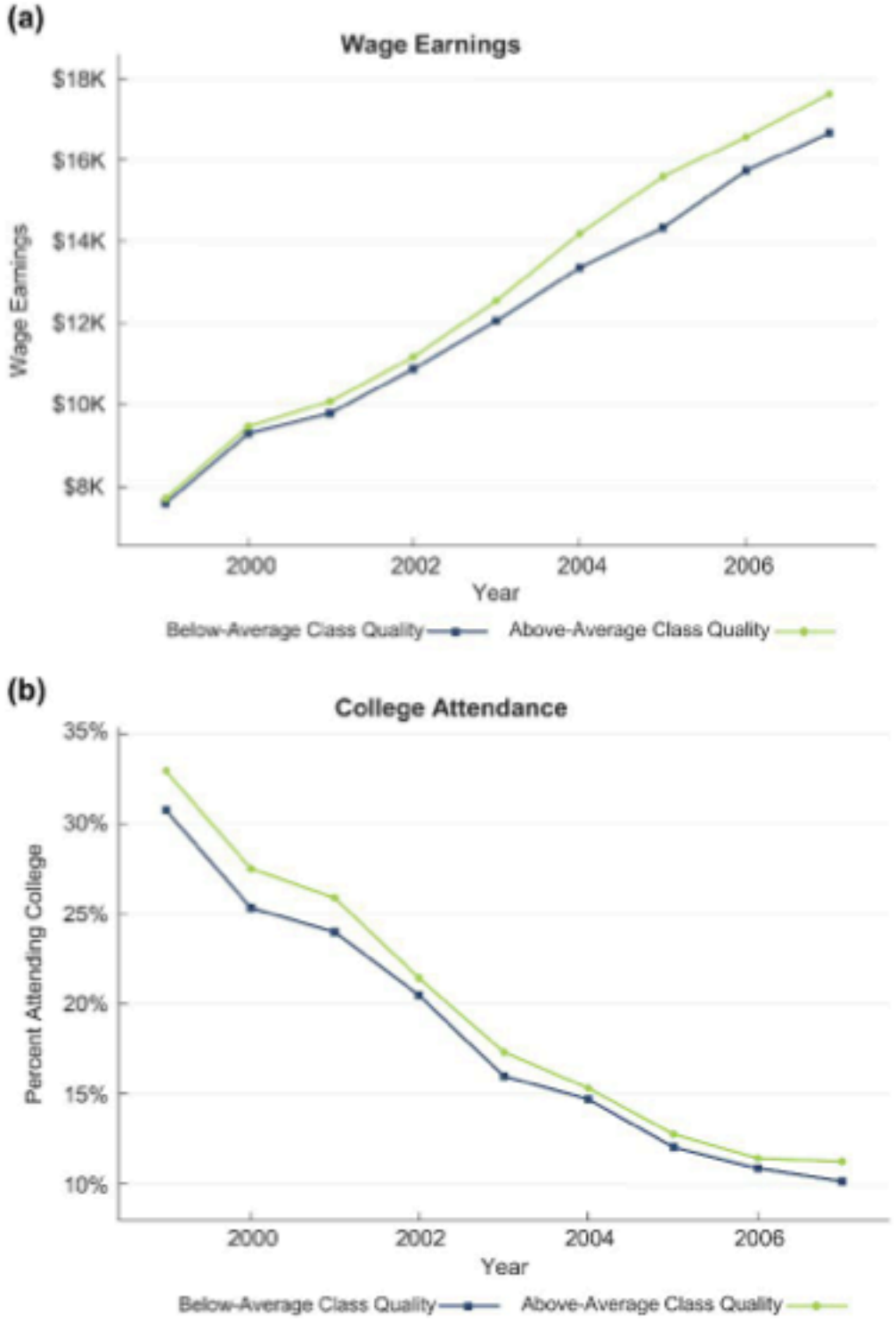

(a)
Wage Earnings
$18K
$16K
$14K
$12K
$10K
$8K
Wage Earnings
2000
2002
2004
2006
Year
Below-Average Class Quality
Above-Average Class Quality
(b)
College Attendance
35%
30%
25%
20%
15%
10%
Percent Attending College
2000
2002
2004
2006
Year
Below-Average Class Quality
Above-Average Class Quality

Figure 2.5

Figure 2.5 (Continued)

Effects of Class Quality by Year

These figures show college attendance rates and mean wage earnings by year (from ages 19 to 27) for students in two groups of classes: those that were above the class quality median and those that were below. Class quality is defined as the difference between the mean end-of-entry-grade test scores of a student's classmates and (grade-specific) schoolmates. Class quality is defined based on the first, randomly assigned STAR classroom (i.e., KG classroom for KG entrants, 1st grade classroom for 1st grade entrants, etc.). Both panels adjust for school-by-entry-grade effects to isolate the random variation in class quality using the procedure in Figure IIc. See notes to Figure I for definitions of wage earnings and college attendance.

Table 2.8 quantifies the impacts of class quality on wage earnings using regressions with the standard vector of parent and student controls used above. Column 1 shows that conditional on the demographic characteristics, a one percentile point increase in class quality increases a student's own test score by 0.66 percentile points. This effect is very precisely estimated, with a t statistic of 27.6, because the intra-class correlation of test scores among students in very large. Column 2 of Table 2.8 shows the effect of class quality on earnings.[73] Conditional on demographic characteristics, a one percentile point increase in class quality increases earnings (averaged from 2005 to 2007) by \$50.6 per year, with a t statistic of 2.9 ($p < 0.01$). To interpret the magnitude of this effect, note that a one standard deviation increase in class quality as measured by peer scores leads to a \$455 (2.9%) increase in earnings at age 27.[74]

[73] Panel C of Online Appendix Table IX replicates the specification in Column 2 to show that class quality has positive impacts on all five alternative measures of wage earnings described above.

[74] Part of the impact of being randomly assigned to a higher quality class in grade w may come from being placed in higher quality classes in subsequent grades. A 1 percentile increase in KG class quality (peer scores) is associated with a 0.15 percentile increase in class quality (peer scores) in grade 1. The analogous effect of grade 1 class quality on grade 2 class quality is 0.37 percentiles.

Table 2.8

EFFECTS OF CLASS QUALITY ON WAGE EARNINGS

Dependent variable	(1) Test score(%)	(2)	(3) Wage earnings ($)	(4)	(5)	(6) College in 2000 (%)	(7) College by age 27 (%)	(8) College quality ($)	(9) Summary index (% of SD)
Class quality	0.662	50.61	61.31	53.44	47.70	0.096	0.108	9.328	0.250
(peer scores)	(0.024)	(17.45)	(20.21)	(24.84)	(18.63)	(0.046)	(0.053)	(4.573)	(0.098)
Entry grade	All	All	All	KG	Grade ≥ 1	All	All	All	All
Observable class chars.			x						
Observations	9,939	10,959	10,859	6,025	4,934	10,959	10,959	10,959	10,959

Notes. Each column reports coefficients from an OLS regression, with standard errors clustered by school in parentheses. Class quality is measured as the difference (in percentiles) between mean end-of-year test scores of the student's classmates and (grade-specific) schoolmates. Class quality is defined based on the first, randomly assigned STAR classroom (i.e. KG class for KG entrants, first-grade class for first-grade entrants, etc.). All specifications control for school-by-entry-grade fixed effects and the vector of demographic characteristics used first in Table IV: a quartic in parent's household income interacted with an indicator for whether the filing parent is ever married between 1996 and 2008, mother's age at child's birth, indicators for parent's 401(k) savings and home ownership, and student's race, gender, free-lunch status, and age at kindergarten. Column (3) includes controls for observable classroom characteristics as in column (6) of Table VII. Column (4) restricts the sample to kindergarten entrants; Column (5) includes only those who enter in grades 1–3. Test score is the average math and reading test score at the end of the year in which the student enters STAR (measured in percentiles). Wage earnings is the individual's mean wage earnings over years 2005–2007 (including 0s for people with no wage earnings). College attendance is measured by receipt of a 1098-T form, issued by higher education institutions to report tuition payments or scholarships. The earnings-based index of college quality is a measure of the mean earnings of all former attendees of each college in the U.S. population at age 28, as described in the text. For individuals who did not attend college, college quality is defined by mean earnings at age 28 of those who did not attend college in the U.S. population. Summary index is the standardized sum of five measures, each standardized on its own before the sum: home ownership, 401(k) retirement savings, marital status, cross-state mobility, and percent of college graduates in the individual's 2007 ZIP code of residence.

The impact of class quality on earnings is estimated much more precisely than the impacts of observable characteristics on earnings because class quality varies substantially across classrooms. Recall from Table 2.5 that students assigned to small classes scored 4.8 percentile points higher on end-of-year tests. If class quality varied only from -2.4 to 2.4, we would be unable to determine whether the relationship between class quality and earnings is significant, as can be seen in Figure 2.4c. By pooling all observable and unobservable sources of variation across classrooms, we obtain more precise (though less policy relevant) estimates of the impact of classroom environments on adult outcomes.

Column 3 of Table 2.8 isolates the variation in class quality that is orthogonal to observable classroom characteristics by controlling for class size, teacher characteristics, and peer characteristics as in column 6 of Table 2.7. Class quality continues to have a significant impact on earnings conditional on these observables, confirming that components of class quality orthogonal to observables matter for earnings.

The preceding specifications pool grades K-3. Column 4 restricts the sample to kindergarten entrants and shows that a one percentile increase in KG class quality raises earnings by \$53.4. Column 5 includes only those who entered STAR after kindergarten. This point estimate is similar to that in column 4, showing that class quality in grades 1-3 matters as much for earnings as class quality in kindergarten.

Columns 6-9 show the impacts of class quality on other adult outcomes. These columns replicate the baseline specification for the full sample in column 2. Columns 6 and 7 show that a 1 percentile improvement in class quality raises college attendance rates by 0.1 percentage points, both at age 20 and before age 27 ($p < 0.05$). Column 8 shows that a one percentile increase in class quality generates an \$9.3 increase in the college quality index ($p < 0.05$). Finally, column 9 shows that a one percentile point improvement in class quality leads to an improvement of 0.25% of a standard deviation in our outcome summary index ($p < 0.05$). Online Appendix Table X reports the impacts of class quality on each of the five outcomes separately and shows that the point estimates of the impacts are positive for all of the outcomes. Online Appendix Table XI documents the heterogeneity of class quality impacts across subgroups. The point estimates of the impacts of class quality are positive for all the groups and outcomes.

Finally, we check the robustness of our results by implementing instrumental-variable methods

of detecting covariance between class effects on scores and earnings. The effects of class quality on test scores and earnings in columns 1 and 2 of Table 2.8 can be combined to produce a jackknife IV estimate of the earnings gain associated with an increase in test scores: \$50.61/0.662 = \$76.48. That is, class-level factors that raise test scores by one percentile point raise earnings by \$76.48 on average. In Online Appendix Table XIII, we show that other IV estimators yield very similar estimates.

While class effects on scores and earnings are highly correlated, a substantial portion of class effects on earnings is orthogonal to our measure of class quality. Using a random effects estimator as in Column 4 of Table 2.7, we find that the standard deviation of class effects on earnings falls from \$1520 to \$1372 after we control for our peer-score class quality measure Δs^{-i}_{cnw}. Hence, roughly $1 - (\frac{1372}{1520})^2 \approx 1/5$ of the variance of the class effect on earnings comes through class effects on test scores.

II.F Fade-Out, Re-Emergence, and Non-Cognitive Skills

In this section, we explore why the impacts of class size and class quality in early childhood fade out on tests administered in later grades but re-emerge in adulthood. In order to have a fixed benchmark to document fade-out, we use only kindergarten entrants throughout this section and analyze the impacts of KG class quality on test scores and other outcomes in later grades.

We first document the fade-out effect using the class quality measure by estimating equation (21) with test scores in each grade as the dependent variable and with the standard vector of parent and student controls as well as school fixed effects. Figure 2.6a plots the estimated impacts on test scores in grades K-8 of increasing KG class quality by one (within-school) standard deviation. A one (within school) SD increase in KG class quality increases end-of-kindergarten test scores by 6.27 percentiles, consistent with our findings above. In grade 1, students who were in a 1 SD better KG class score approximately 1.50 percentile points higher on end-of-year tests, an effect that is significant with $p < 0.001$. The effect gradually fades over time, and by grade 4 students who were in a better KG class no longer score significantly higher on tests.[75]

[75] This fade-out effect is consistent with the rapid fade-out of teacher effects documented by Jacob, Lefgren, and Sims (2008), Kane and Staiger (2008), and others.

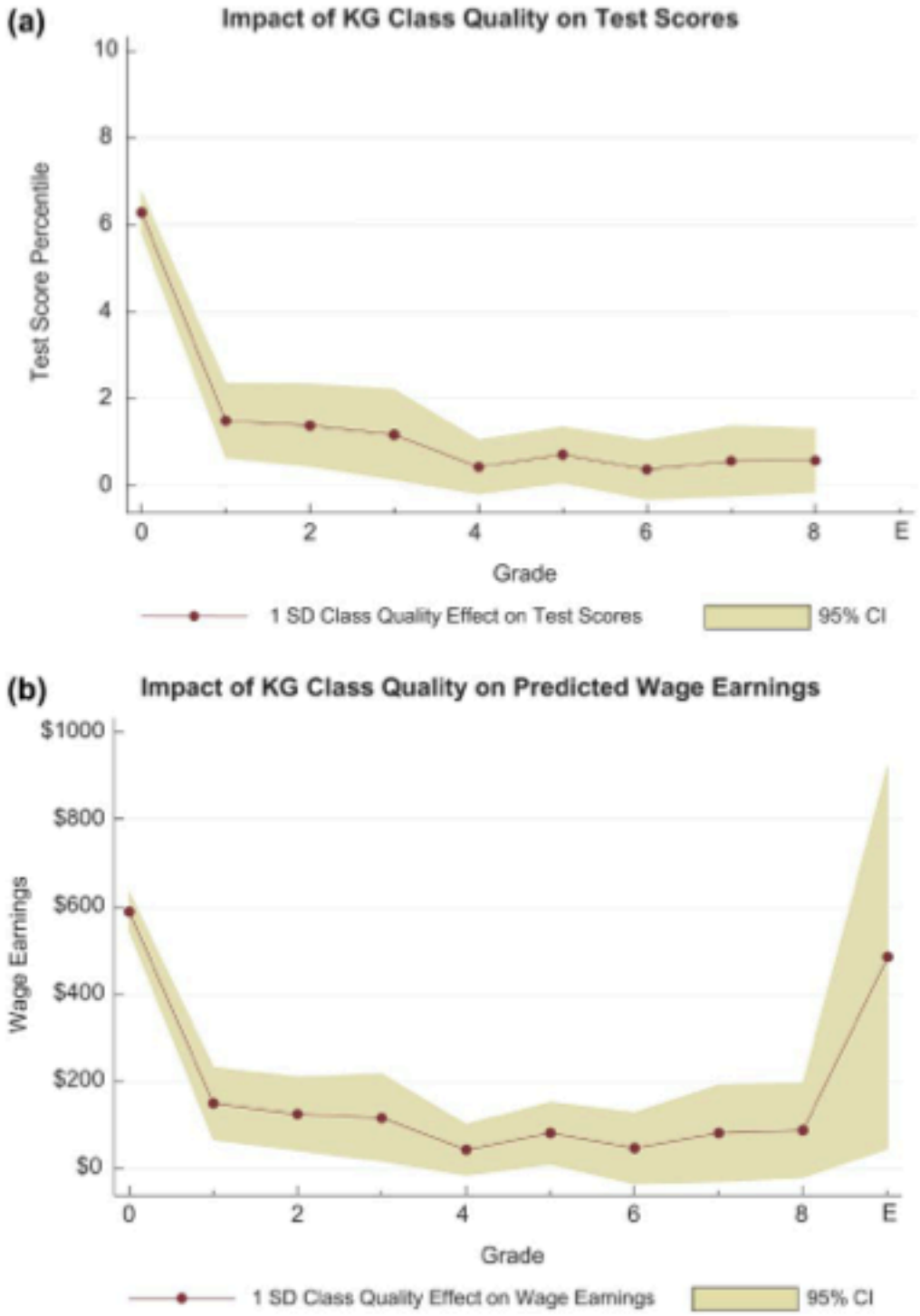
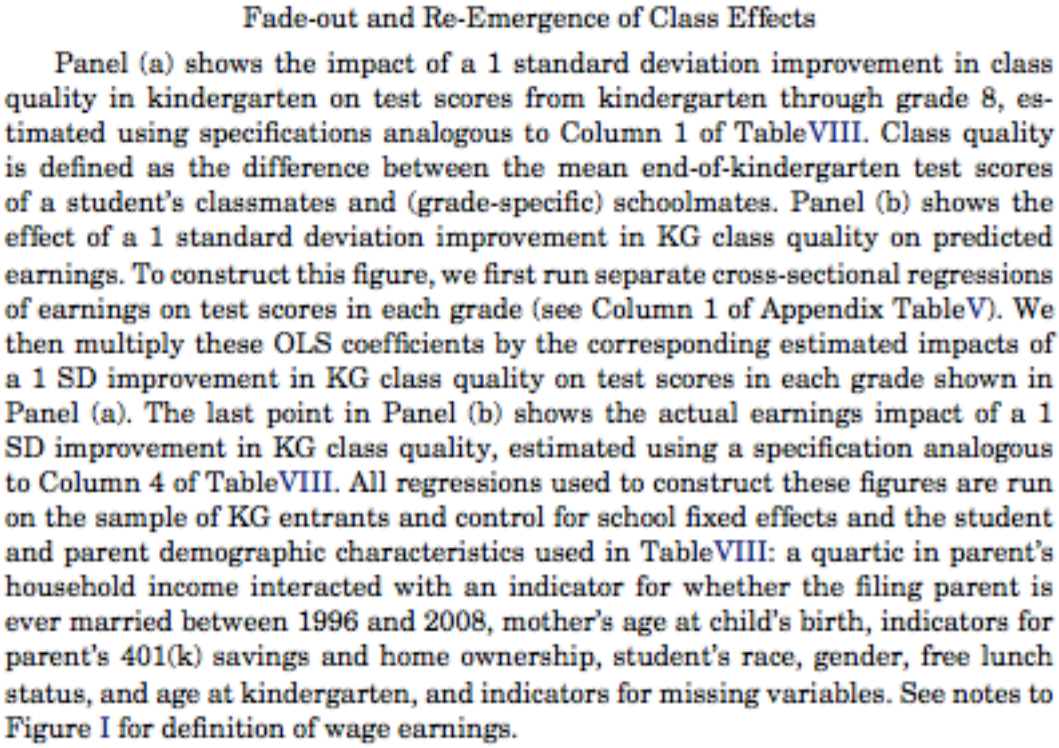

(a)
Impact of KG Class Quality on Test Scores
Test Score Percentile
10
8
6
4
2
0
0
2
4
6
8
E
Grade
1 SD Class Quality Effect on Test Scores
95% CI
(b)
Impact of KG Class Quality on Predicted Wage Earnings
Wage Earnings
$1000
$800
$600
$400
$200
$0
0
2
4
6
8
E
Grade
1 SD Class Quality Effect on Wage Earnings
95% CI

Figure 2.6

Figure 2.6 (Continued)



Fade-out and Re-Emergence of Class Effects

Panel (a) shows the impact of a 1 standard deviation improvement in class quality in kindergarten on test scores from kindergarten through grade 8, estimated using specifications analogous to Column 1 of TableVIII. Class quality is defined as the difference between the mean end-of-kindergarten test scores of a student's classmates and (grade-specific) schoolmates. Panel (b) shows the effect of a 1 standard deviation improvement in KG class quality on predicted earnings. To construct this figure, we first run separate cross-sectional regressions of earnings on test scores in each grade (see Column 1 of Appendix TableV). We then multiply these OLS coefficients by the corresponding estimated impacts of a 1 SD improvement in KG class quality on test scores in each grade shown in Panel (a). The last point in Panel (b) shows the actual earnings impact of a 1 SD improvement in KG class quality, estimated using a specification analogous to Column 4 of TableVIII. All regressions used to construct these figures are run on the sample of KG entrants and control for school fixed effects and the student and parent demographic characteristics used in TableVIII: a quartic in parent's household income interacted with an indicator for whether the filing parent is ever married between 1996 and 2008, mother's age at child's birth, indicators for parent's 401(k) savings and home ownership, student's race, gender, free lunch status, and age at kindergarten, and indicators for missing variables. See notes to Figure I for definition of wage earnings.

If a one percentile increase in 8th grade test scores is more valuable than a one percentile increase in KG test scores, then the evidence in Figure 2.6a would not necessarily imply that the effects of early childhood education fade out. To evaluate this possibility, we convert the test score impacts to predicted earnings gains. We run separate OLS regressions of earnings on the test scores for each grade from K-8 to estimate the cross-sectional relationship between each grade's test score and earnings (see Online Appendix Table V, Column 1 for these coefficients). We then multiply the class quality effect on scores shown in Figure 2.6a by the corresponding coefficient on scores from the OLS earnings regression. Figure 2.6b plots the earnings impacts predicted by the test score gains in each grade that arise from attending a better KG class. The pattern in Figure 2.6b looks very similar to that in Figure 2.6a, showing that there is indeed substantial fade-out of the KG class quality effect on predicted earnings. By 4th grade, one would predict less than a $50 per year gain in earnings from a better KG class based on observed test score impacts.

The final point in Figure 2.6b shows the actual observed earnings impact of a one SD improvement in KG class quality. The actual impact of $483 is similar to what one would have predicted based on the improvement in KG test scores ($588). The impacts of early childhood education re-emerge in adulthood despite fading out on test scores in later grades.

Non-Cognitive Skills. One potential explanation for fade-out and re-emergence is the acquisition of non-cognitive skills (e.g. Heckman 2000, Heckman, Stixrud, and Urzua 2006, Lindqvist and Vestman 2011). We evaluate whether non-cognitive skills could explain our findings using data on non-cognitive measures collected for a subset of STAR students in grades 4 and 8.[76]

Finn et al. (2007) and Dee and West (2008) describe the non-cognitive measures in the STAR data in detail; we provide a brief summary here. In grade 4, teachers in the STAR schools were asked to evaluate a random subset of their students on a scale of 1-5 on several behavioral measures, such as whether the student "annoys others." These responses were consolidated into four standardized scales measuring each student's effort, initiative, nonparticipatory behavior, and how the student is seen to "value" the class. In grade 8, math and English teachers were asked to rate a subset of their students on a similar set of questions, which were again consolidated into the same four

[76] Previous studies have used the STAR data to investigate whether class size affects non-cognitive skills (Finn et al. 1989, Dee and West 2008). They find mixed evidence on the impact of class size on non-cognitive skills: statistically significant impacts are detected in grade 4, but not in grade 8. Here, we analyze the impacts of our broader class quality measure.

standardized scales. To obtain a measure analogous to our percentile measure of test scores, we construct percentile measures for these four scales and compute the average percentile score for each student. For 8th grade, we then take the average of the math and English teacher ratings.

Among the 6,025 students who entered Project STAR in KG and whom we match in the IRS data, we have data on non-cognitive skills for 1,671 (28%) in grade 4 and 1,780 (30%) in grade 8. The availability of non-cognitive measures for only a subset of the students who could be tracked until grade 8 naturally raises concerns about selective attrition. Dee and West (2008) investigate this issue in detail, and we replicate their findings with our expanded set of parental characteristics. In grade 8, we find no significant differences in the probability of having non-cognitive data by KG classrooms or class types (small vs. large), and confirm that in this sample the observable background characteristics are balanced across classrooms and class types. In grade 4, non-cognitive data are significantly more likely to be available for students assigned to small classes, but among the sample with non-cognitive data there are no significant differences in background characteristics across classrooms or class types. Hence, the sample for whom we have non-cognitive data appear to be balanced across classrooms at least on observable characteristics.

We begin by estimating the cross-sectional correlation between non-cognitive outcomes and earnings. Column 1 of Table 2.9 shows that a 1 percentile improvement in non-cognitive measures in grade 4 is associated with a $106 gain in earnings conditional on the standard vector of demographic characteristics used above and school-by-entry-grade fixed effects. Column 2 shows that controlling for math and reading test scores in grade 4 reduces the predictive power of non-cognitive scores only slightly, to $88 per percentile. In contrast, column 3 shows that non-cognitive skills in grade 4 are relatively weak predictors of 8th grade test scores when compared with math and reading scores in 4th grade. Because non-cognitive skills appear to be correlated with earnings through channels that are not picked up by subsequent standardized tests, they could explain fade-out and re-emergence.

Table 2.9

EFFECTS OF KG CLASS QUALITY ON NONCOGNITIVE SKILLS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			Grade 8	Grade 4				Grade 8	
Dependent variable:	Wage reading ($)		Math + reading (%)	Math + reading (%)	Noncog (%)	Math + reading (%)	Noncog (%)	Math + reading (%)	Noncog (%)
Grade 4 noncog. score	106	87.7	0.059						
	(16.0)	(20.4)	(0.017)						
Grade 4 math + reading score		36.4	0.671						
		(24.7)	(0.023)						
Class quality (peer scores)				0.047	0.153	0.064	0.128		
				(0.035)	(0.065)	(0.041)	(0.054)		
Teacher with >10 years experience								0.292	2.60
								(0.878)	(1.41)
Observations	1,671	1,360	1,254	4,023	1,671	4,448	1,780	4,432	1,772

Notes. Each column reports coefficients from an OLS regression, with standard errors clustered by school in parentheses. All specifications include only the subsample of students who entered a STAR school in kindergarten, and control for school fixed effects and the vector of demographic characteristics used first in TableIV: a quartic in parent's household income interacted with an indicator for whether the filing parent is ever married between 1996 and 2008, mother's age at child's birth, indicators for parent's 401(k) savings and home ownership, and student's race, gender, free-lunch status, and age at kindergarten. Wage earnings is the individual's mean wage earnings over years 2005–2007 (including 0s for people with no wage earnings). Grades 4 and 8 noncognitive scores are based on teacher surveys of student behavior across four areas: effort, initiative, engagement in class, and whether the student values school. We average the four component scores and convert them into within-sample percentile ranks. Math + reading scores are average math and reading test scores (measured in percentiles) at the end of the relevant year. Class quality is measured as the difference (in percentiles) between mean end-of-year test scores of the student's classmates and (grade-specific) schoolmates in kindergarten. Teacher experience is the number of years the KG teacher taught at any school before the student's year of entry into a STAR school.

To further evaluate this mechanism, we investigate the effects of KG class quality on non-cognitive skills in grade 4 and 8. As a reference, column 4 shows that a 1 percentile improvement in KG class quality increases a student's test scores in grade 4 by a statistically insignificant 0.05 percentiles. In contrast, column 5 shows that the same improvement in KG class quality generates a statistically significant increase of 0.15 percentiles in the index of non-cognitive measures in grade 4. Columns 6 and 7 replicate columns 4 and 5 for grade 8.[77] Again, KG class quality does not have a significant impact on 8th grade test scores but has a significant impact on non-cognitive measures. Finally, columns 8 and 9 show that the experience of the student's teacher in kindergarten – which we showed above also impacts earnings – has a small and statistically insignificant impact on test scores but a substantially larger impact on non-cognitive measures in 8th grade ($p = 0.07$).[78]

We can translate the impacts on non-cognitive skills into predicted impacts on earnings following the method in Figure 2.6b. We regress earnings on the non-cognitive measure in grade 4, conditioning on demographic characteristics, and obtain an OLS coefficient of \$101 per percentile. Multiplying this OLS coefficient by the estimated impact of class quality on non-cognitive skills in grade 4, we predict that a 1 SD improvement in KG class quality will increase earnings by \$139. The same exercise for 4th grade math+reading test scores yields a predicted earnings gain of \$40. These results suggest that improvements in non-cognitive skills explain a larger share of actual earnings gains than improvements in cognitive performance, consistent with Heckman et al.'s (2010) findings for the Perry Preschool program. In contrast, a one standard deviation increase in class quality is predicted to raise 8th grade test scores by only 0.47 percentiles based on its observed impacts on non-cognitive skills in grade 4 and the cross-sectional correlation between grade 4 non-cognitive skills and grade 8 test scores. This predicted impact is quite close to the actual impact of class quality on 8th grade scores of 0.57 percentiles. Hence, the impacts of class quality on non-cognitive skills is consistent with both fade-out on scores and re-emergence on adult outcomes.

Intuitively, a better kindergarten classroom might simultaneously increase performance on end-of-year tests and improve untested non-cognitive skills. For instance, a KG teacher who is able to

[77] We use all KG entrants for whom test scores are available in columns 4 and 6 to increase precision. The point estimates on test score impacts are similar for the subsample of students for whom non-cognitive data are available.

[78] Online Appendix Table XIV decomposes the relationships described in Table IX into the four constituent components of non-cognitive skill.

make her students memorize vocabulary words may instill social skills in the process of managing her classroom successfully. These non-cognitive skills may not be well measured by standardized tests, leading to very rapid fade-out immediately after KG. However, these skills could still have returns in the labor market.

Although non-cognitive skills provide one plausible explanation of the data, our analysis is far from definitive proof of the importance of non-cognitive skills. The estimates of non-cognitive impacts could suffer from attrition bias and are somewhat imprecisely estimated. Moreover, our analysis does not show that manipulating non-cognitive skills directly has causal impacts on adult outcomes. We have shown that high quality KG classes improve both non-cognitive skills and adult outcomes, but the mechanism through which adult outcomes are improved could run through another channel that is correlated with the acquisition of non-cognitive skills. It would be valuable to analyze interventions that target non-cognitive skills directly in future work.

II.G Conclusion

The impacts of education have traditionally been measured by achievement on standardized tests. This paper has shown that the classroom environments that raise test scores also improve long-term outcomes. Students who were randomly assigned to higher quality classrooms in grades K-3 earn more, are more likely to attend college, save more for retirement, and live in better neighborhoods. Yet the same students do not do much better on standardized tests in later grades. These results suggest that policy makers may wish to rethink the objective of raising test scores and evaluating interventions via long-term test score gains. Researchers who had examined only the impacts of STAR on test scores would have incorrectly concluded that early childhood education does not have long-lasting impacts. While the quality of education is best judged by directly measuring its impacts on adult outcomes, our analysis suggests that *contemporaneous* (end-of-year) test scores are a reasonably good short-run measure of the quality of a classroom.

We conclude by using our empirical estimates to provide rough calculations of the benefits of various policy interventions (see Online Appendix C for details). These cost-benefit calculations rely on several strong assumptions. We assume that the percentage gain in earnings observed at age 27 remains constant over the lifecycle. We ignore non-monetary returns to education (such as reduced crime) as well as general equilibrium effects. We discount earnings gains at a 3% annual

rate back to age 6, the point of the intervention.

(1) Class Quality. The random-effects estimate reported in column 4 of Table 2.7 implies that increasing class quality by one standard deviation of the distribution within schools raises earnings by $1,520 (9.6%) at age 27. Under the preceding assumptions, this translates into a lifetime earnings gain of approximately $39,100 for the average individual. For a classroom of twenty students, this implies a present-value benefit of $782,000 for improving class quality for a single year by one (within-school) standard deviation. This large figure includes all potential benefits from an improved classroom environment, including better peers, teachers, and random shocks, and hence is useful primarily for understanding the stakes at play in early childhood education. It is less helpful from a policy perspective because one cannot implement interventions that directly improve classroom quality. This motivates the analysis of class size and better teachers, two factors that contribute to classroom quality.

(2) Class Size. We calculate the benefits of reducing class size by 33% in two ways. The first method uses the estimated earnings gain from being assigned to a small class reported in column 5 of Table 2.5. The point estimate of $4 in Table V translates into a lifetime earnings gain from reducing class size by 33% for one year of $103 in present value per student, or $2,057 for a class that originally had twenty students. But this estimate is imprecise: the 95% confidence interval for the lifetime earnings gain of reducing class size by 33% for one year ranges from -$17,500 to $17,700 per child. To obtain more precision, we predict the benefits of class size reduction using the estimated impact of classroom quality on scores and earnings. We estimate that a 1 percentile increase in class quality raises test scores by 0.66 percentiles and earnings by $50.6, implying an earnings gain of $76.7 per percentile increase in test scores. Next, we make the strong assumption that the ratio of earnings gains to test score gains is the same for changes in class size as it is for improvements in class quality more generally. Under this assumption, a 33% class size reduction in grades K-3 (which raised test scores by 4.8 percentiles) is predicted to raise earnings by $4.8 \times \$76.7 = \$368$ (2.3%) at age 27. This calculation implies a present value earnings gain from class size reduction of $9,460 per student and $189,000 for the classroom.[79]

(3) Teachers. We cannot directly estimate the total impacts of teachers on earnings in this study

[79] Krueger (1999) projects a gain from small-class attendence of $\$9,603$ for men and $\$7,851$ for women. Neither of our estimates are statistically distinguishable from these predictions.

because we observe each teacher in only one classroom, making it impossible to separate teacher effects from peer effects and classroom-level shocks. However, we can predict the magnitudes of teacher effects as measured by value-added on test scores by drawing upon prior work. Rockoff (2004), Rivkin, Hanushek, and Kain (2005), and Kane and Staiger (2008) use datasets with multiple classrooms per teacher to estimate that a one standard deviation increase in teacher quality raises test scores by between 0.1 and 0.2 standard deviations (2.7-5.4 percentiles).[80] Under the strong assumption that the ratio of earnings gains to test score gains is the same for changes in teacher quality and class quality more broadly, this test score gain implies an earnings gain of $208-$416 (1.3%-2.6%) at age 27 and a present-value earnings gain ranging from $5,350-$10,700 per student. Hence, we predict that a one standard deviation improvement in teacher quality in a single year would generate earnings gains between $107,000 and $214,000 for a classroom of twenty students. These predictions are roughly consistent with the findings of Chetty, Friedman, and Rockoff (2011), who directly estimate the impacts of teacher value-added on earnings using a dataset that contains information on multiple classrooms per teacher.

Our results suggest that good teachers could potentially create great social value, perhaps several times larger than current teacher salaries.[81] However, our findings do not have direct implications for optimal teacher salaries or merit pay policies as we do not know whether higher salaries or merit pay would improve teacher quality.[82] Relative to efforts that seek to improve the quality of teachers, class size reductions have the important advantage of being more well-defined and straightforward to implement. However, reductions in class size must be implemented carefully to generate improvements in outcomes. If schools are forced to reduce teacher and class quality along other dimensions when reducing class size, the net gains from class size reduction may be diminished (Jepsen and Rivkin 2009, Sims 2009).

Finally, our analysis raises the possibility that differences in school quality perpetuate income inequality. In the U.S., higher income families have access to better public schools on average

[80] We use estimates of the impacts of teacher quality on scores from other studies to predict earnings gains because we do not have repeat observations on teachers in our data. In future work, it would be extremely valuable to link datasets with repeat observations on teachers to administrative data on students in order to measure teachers' impacts on earnings directly.

[81] According to calculations from the 2006-2008 American Community Survey, the mean salary for elementary and middle school teachers in the U.S. was $39,164 (in 2009 dollars).

[82] An analogy with executive compensation might be helpful in understanding this point. CEOs' decisions have large impacts on the firms they run, and hence can create or destroy large amounts of economic value. But this does not necessarily imply that increasing CEO compensation or pay-for-performance would improve CEO decisions.

because of property-tax finance. Using the class quality impacts reported above, Chetty and Friedman (2011) estimate that the intergenerational correlation of income would fall by roughly 1/3 if all children attended schools of the same quality. Improving early childhood education in disadvantaged areas – e.g. through federal tax credits or tax policy reforms – could potentially reduce inequality in the long run.

Appendix A: Algorithm for Matching STAR Records to Tax Data

The tax data were accessed through contract TIRNO-09-R-00007 with the Statistics of Income (SOI) Division at the US Internal Revenue Service. Requests for research contracts by SOI are posted online at the Federal Business Opportunities https://www.fbo.gov/. SOI also welcomes research partnerships between outside academics and internal researchers at SOI.

STAR records were matched to tax data using social security number (SSN), date of birth, gender, name, and STAR elementary school ZIP code. Note that STAR records do not contain all the same information. Almost every STAR record contains date of birth, gender, and last name. Some records contain no SSN while others contain multiple possible SSNs. Some records contain no first name. A missing field yielded a non-match unless otherwise specified.

We first discuss the general logic of the match algorithm and then document the routines in detail. The match algorithm was designed to match as many records as possible using variables that are *not* contingent on ex post outcomes. SSN, date of birth, gender, and last name in the tax data are populated by the Social Security Administration using information that is not contingent on ex post outcomes. First name and ZIP code in tax data are contingent on observing some ex post outcome. First name data derive from information returns, which are typically generated after an adult outcome like employment (W-2 forms), college attendance (1098-T forms), and mortgage interest payment (1098 forms). The ZIP code on the claiming parent's 1040 return is typically from 1996 and is thus contingent on the ex post outcome of the STAR subject not having moved far from her elementary school by age 16.

89.8% of STAR records were matched using only ex ante information. The algorithm first matched as many records as possible using only SSN, date of birth, gender, and last name. It then used first name only to *exclude* candidate matches based on date of birth, gender, and last name, often leaving only one candidate record remaining. Because that exclusion did not condition on an information return having been filed on behalf of that remaining candidate, these matches also did not condition on ex post outcomes.

The match algorithm proceeded as follows, generating seven match types denoted A through G. The matches generated purely through ex-ante information are denoted A through E below and account for 89.8% of STAR records. Matches based on ex-post-information are denoted F and

G below and constitute an additional 5.4% of STAR records. The paper reports results using the full 95.0% matched sample, but all the qualitative results hold in the 89.8% sample matched using only ex ante information.

1. Match STAR records to tax records by SSN. For STAR records with multiple possible SSNs, match on all of these SSNs to obtain a set of candidate tax record matches for each STAR record with SSN information. Each candidate tax record contains date of birth, gender, and first four letters of every last name ever assigned to the SSN.

    - Match Type A. Keep unique matches after matching on first four letters of last name, date of birth, and gender.
    - Match Type B. Refine non-unique matches by matching on either first four letters of last name or on "fuzzy" date of birth. Then keep unique matches. Fuzzy date of birth requires the absolute value of the difference between STAR record and tax record dates of birth to be in the set {0,1,2,3,4,5,9,10,18,27} in days, in the set {1,2} in months, or in the set {1} in years. These sets were chosen to reflect common mistakes in recorded dates of birth, such as being off by one day (e.g. 12 vs. 13) or inversion of digits (e.g. 12 vs. 21).

2. Match residual unmatched STAR records to tax records by first four letters of last name, date of birth, and gender.

    - Match Type C. Keep unique matches.
    - Match Type D. Refine non-unique matches by excluding candidates who have a first name issued on information returns (e.g. W-2 forms, 1098-T forms, and various 1099 forms) that does not match the STAR first name on first four letters when the STAR first name is available. Then keep unique matches.
    - Match Type E. Refine residual non-unique matches by excluding candidates who have SSNs that, based on SSN area number, were issued from outside the STAR region (Tennessee and neighboring environs). Then keep unique matches.
    - Match Type F. Refine residual non-unique matches by keeping unique matches after each of the following additional criteria is applied: require a first name match when STAR

first name is available, require the candidate tax record's SSN to have been issued from the STAR region, and require the first three digits of the STAR elementary school ZIP code to match the first three digits of the ZIP code on the earliest 1040 return on which the candidate tax record was claimed as a dependent.

3. Match residual unmatched STAR records to tax records by first four letters of last name and fuzzy date of birth.

   - Match Type G. Keep unique matches after each of several criteria is sequentially applied. These criteria include matches on first name, last name, and middle initial using the candidate tax record's information returns; on STAR region using the candidate tax record's SSN area number; and between STAR elementary school ZIP code and ZIP code on the earliest 1040 return on which the candidate tax record was claimed as a dependent.

The seven match types cumulatively yielded a 95.0% match rate:

Match type	Frequency	Percent	Cumulative percent
A	7036	60.8%	60.8%
B	271	2.3%	63.1%
C	699	6.0%	69.2%
D	1391	12.0%	81.2%
E	992	8.6%	89.8%
F	299	2.6%	92.4%
G	304	2.6%	95.0%

Identifiers such as names and SSN's were used solely for the matching procedure. After the match was completed, the data were de-identified (i.e., individual identifiers such as names and SSNs were stripped) and the statistical analysis was conducted using the de-identified dataset.

Appendix B: Derivations for Measurement of Unobserved Class Quality

This appendix derives the estimators discussed in the empirical model in Section V and quantifies the degree of attenuation and reflection bias. We first use equations (14) and (15) to define

average of test scores and earnings within each class c and school n:

$$\begin{aligned} s_{cn} &= d_n + z_{cn} + a_{cn} \\ y_{cn} &= \delta_n + \beta z_{cn} + z^Y_{cn} + \rho a_{cn} + \nu_{cn} \\ s_n &= d_n + z_n + a_n \\ y_n &= \delta_n + \beta z_n + z^Y_n + \rho a_n + \nu_n. \end{aligned}$$

We define variables demeaned within schools as

$$\begin{aligned} s_{icn} - s_n &= z_{cn} - z_n + a_{icn} - a_n \\ \Delta s_{cn} \equiv s_{cn} - s_n &= z_{cn} - z_n + a_{cn} - a_n, \\ y_{icn} - y_n &= \beta(z_{cn} - z_n) + (z^Y_{icn} - z^Y_n) + \rho(a_{icn} - a_n) + \nu_{icn} - \nu_n \\ y_{cn} - y_n &= \beta(z_{cn} - z_n) + (z^Y_{cn} - z^Y_n) + \rho(a_{cn} - a_n) + \nu_{cn} - \nu_n. \end{aligned}$$

Recall that a_{icn} and ν_{icn} are independent of each other and z_{cn}. Let $\sigma^2 = var(a_{icn})$. We assume in parts 1 and 2 below that $z_{cn}, z^Y_{cn} \perp a_{icn}$, ruling out peer effects. Note also that, as $z_{cn} \perp z^Y_{cn}$, the component of classroom environments that affects only test scores drops out entirely of the covariance analysis below. In what follows, we take the number of students per class I and the number of classrooms per school C as fixed and analyze the asymptotic properties of various estimators as the number of schools $N \to \infty$.

1. Mean score estimator. The simplest proxy for class quality is the average test score within a class. Since we include school fixed effects in all specifications, s_{cn} is equivalent to Δs_{cn} as defined above. Therefore, consider the following (school) fixed effects OLS regression:

$$y_{icn} = \alpha_n + b^M \Delta s_{cn} + \varepsilon_{icn}. \tag{22}$$

As the number of schools $N \to \infty$, the coefficient estimate $\hat{b}^M$ converges to

$$plim_{\ N \longrightarrow \infty}\ \hat{b}^M = \frac{cov(y_{icn} - y_n, s_{cn} - s_n)}{var(s_{cn} - s_n)},$$

which we can rewrite as

$$\begin{aligned} plim\ _{N\longrightarrow\infty}\ \hat{b}^M &= \frac{cov\left(\beta(z_{cn}-z_n)+\rho(a_{icn}-\frac{\sum_k\sum_j a_{jkn}}{I\cdot C}), z_{cn}-z_n+\frac{\sum_j a_{jcn}}{I}-\frac{\sum_k\sum_j a_{jkn}}{I\cdot C}\right)}{var\left(z_{cn}-z_n+\frac{\sum_j a_{jcn}}{I}-\frac{\sum_k\sum_j a_{jkn}}{I\cdot C}\right)} \\ &= \frac{\beta var(z_{cn}-z_n)+\rho\sigma^2\frac{C-1}{IC}}{var(z_{cn}-z_n)+\sigma^2\frac{C-1}{IC}}. \end{aligned}$$

Even absent class effects ($\beta = 0$), we obtain $plim\ _{N\longrightarrow\infty}\ \hat{b}^M > 0$ if I is finite and $\rho > 0$. With finite class size, b^M is upward-biased due to the correlation between wages and own-score, which is included within the class quality measure.

2. Leave-out mean estimator. We address the upward bias due to the own observation problem using a leave-out mean estimator. Consider the OLS regression with school fixed effects

$$y_{icn} = \alpha_n + b^{LM}\Delta s_{cn}^{-i} + \varepsilon_{icn}. \tag{23}$$

where $\Delta s_{cn}^{-i} = s_{cn}^{-i} - s_n^{-i}$ is defined as in equation (18). The coefficient b^{LM} converges to

$$plim\ _{N\longrightarrow\infty}\ \hat{b}^{LM} = \frac{cov(y_{icn}-y_n, s_{cn}^{-i}-s_n^{-i})}{var(s_{cn}^{-i}-s_n^{-i})},$$

which we can rewrite as

$$\begin{aligned} plim\ _{N\longrightarrow\infty}\ \hat{b}^{LM} &= \frac{cov\left(\beta(z_{cn}-z_n)+\rho(a_{icn}-a_n), \frac{IC}{IC-1}(z_{cn}-z_n)+\frac{1}{I-1}\sum_{j\neq i}a_{jcn}-\frac{1}{IC-1}\sum_k\sum_{j\neq i}a_{jkn}\right)}{var\left(\frac{IC}{IC-1}(z_{cn}-z_n)+\frac{1}{I-1}\sum_{j\neq i}a_{jcn}-\frac{1}{IC-1}\sum_k\sum_{j\neq i}a_{jkn}\right)} \\ &= \beta\times\frac{\frac{IC}{IC-1}var(z_{cn}-z_n)}{\frac{(IC)^2}{(IC-1)^2}var(z_{cn}-z_n)+\frac{\sigma^2}{I-1}-\frac{\sigma^2}{I\cdot C-1}} \end{aligned}$$

Hence, $plim\ _{N\longrightarrow\infty}\ \hat{b}^{LM} = 0$ if and only if $\beta var(z_{cn}-z_n) = 0$ (no class effects) even when I and C are finite.[83] However, b^{LM} is attenuated relative to β because peer scores are a noisy measure of class quality.

[83] The leave-out mean estimator b^{LM} is consistent as the number of schools grows large, but is downward biased in small samples because own scores s_{icn} and peer scores Δs_{cn}^{-i} are mechanically negatively correlated within classrooms. Monte Carlo simulations suggest that this finite sample bias is negligible in practice with the number of schools and classrooms in the STAR data.

Quantifying the degree of attenuation bias. We can quantify the degree of attenuation bias by using the within-class variance of test scores as an estimate of $\sigma^2 = var(a_{icn})$. First, note that:

$$\begin{aligned}
\widehat{var}(z_{cn} - z_n) &= \frac{(IC-1)^2}{(IC)^2}\left[\widehat{var}\left(s_{cn}^{-i} - s_n^{-i}\right) - \left(\frac{\hat{\sigma}^2}{I-1} - \frac{\hat{\sigma}^2}{I \cdot C - 1}\right)\right] \\
&= \frac{(83.63)^2}{(84.73)^2}\left[81.75 - \left(\frac{437.4}{19.07} - \frac{437.4}{83.63}\right)\right] \\
&= 62.39
\end{aligned}$$

where we use the sample harmonic means for IC, $IC-1$, and $I-1$ because the number of students in each class and school varies across the sample. This implies an estimate of bias of

$$\frac{\frac{83.63}{84.73}62.39}{\frac{(83.63)^2}{(84.73)^2}62.39 + \frac{437.4}{19.07} - \frac{437.4}{83.63}} = 0.773.$$

That is, b^{LM} is attenuated relative to β by 22.7%. Note that this bias calculation assumes that all students in the class were randomly assigned, which is true only in KG. In later grades, the degree of attenuation in b^{LM} when equation (23) is estimated using new entrants is larger than 22.7%, because existing students were not necessarily re-randomized at the start of subsequent grades.

3. Peer effects and reflection bias. With peer effects, the assumption $z_{cn} \perp a_{icn}$ does not hold. We expect z_{cn} and a_{icn} to be positively correlated with peer effects as a higher ability student has a positive impact on the class. This leads to an upward bias in both b^{LM} and b^{SS} due to the reflection problem. To characterize the magnitude of this bias, consider a standard linear-in-the-means model of peer effects, in which

$$z_{cn} = t_{cn} + \frac{\theta}{I}\sum_j a_{jcn}$$

with $t_{cn} \perp a_{jcn}$ for all j. Here t_{cn} represents the component of class effects independent of peer effects (e.g., a pure teacher effect). The parameter $\theta > 0$ captures the strength of peer effects. Averaging across classrooms within a school implies that

$$z_n = t_n + \frac{\theta}{IC}\sum_k\sum_j a_{jkn}.$$

In this model, the leave-out mean proxy of class quality is

$$\Delta s_{cn}^{-i} = s_{cn}^{-i} - s_n^i = \frac{IC}{IC-1}(t_{cn} - t_n) + \theta\frac{IC}{IC-1}(a_{cn} - a_n) + \frac{1}{I-1}\sum_{j\neq i} a_{jcn} - \frac{1}{IC-1}\sum_k \sum_{j\neq i} a_{jkn}$$

and as N grows large $\hat{b}^{LM}$ converges to

$$\begin{aligned} plim\ _{N\longrightarrow\infty}\ \hat{b}^{LM} &= \frac{cov(y_{icn} - y_n, s_{cn}^{-i} - s_n^{-i})}{var(s_{cn}^{-i} - s_n^{-i})} \\ &= \frac{\beta \cdot \left[\frac{IC}{IC-1} var(t_{cn} - t_n) + (\theta + \theta^2)\sigma^2 \frac{C-1}{IC-1}\right] + \rho\theta\sigma^2 \frac{C-1}{IC-1}}{\frac{(IC)^2}{(IC-1)^2} var(t_{cn} - t_n) + (2\theta + \theta^2)\sigma^2 \frac{IC(C-1)}{(IC-1)^2} + \frac{\sigma^2}{I-1} - \frac{\sigma^2}{I\cdot C-1}} \end{aligned}$$

The last term in the numerator is the reflection bias that arises because a high ability student has both high earnings (through ρ) and a positive impact on peers' scores (through θ). Because of this term, we can again obtain $plim\ _{N\longrightarrow\infty}\ \hat{b}^{LM} > 0$ even when $\beta = 0$. This bias occurs iff $\theta > 0$ (i.e., we estimate $b^{LM} > 0$ only if there are peer effects on test scores). This bias is of order $\frac{1}{I}$ since any given student is only one of I students in a class that affects class quality.

Bounding the degree of reflection bias. We use the estimated impact of KG class quality on 8th grade test scores to bound the degree of reflection bias in our estimate of the impact of class quality on earnings. Recall that the reflection bias arises because a high ability student has better long-term outcomes and also has a positive impact on peers' kindergarten test scores. Therefore, the same reflection bias is present when estimating $\hat{b}^{LM}$ using eighth grade test scores as the outcome instead of earnings.

Denote by $\hat{b}_e^{LM}$ the estimated coefficient on Δs_{cn}^{-i} when the outcome y is earnings and $\hat{b}_s^{LM}$ the same coefficient when the outcome y is grade 8 test scores.[84] Similarly, denote by ρ_e and ρ_s the (within class) correlation between individual kindergarten test score and earnings or eighth grade test score. Under our parametric assumptions, these two parameters can be estimated by an OLS regression $y_{icn} = \alpha_{cn} + \rho s_{icn} + \varepsilon_{icn}$ that includes class fixed effects.

To obtain an upper bound on the degree of reflection bias, we make the extreme assumption that the effect of kindergarten class quality on eighth grade test scores ($\hat{b}_s^{LM}$) is due entirely to the

[84] The latest test score we have in our data is in grade 8. We find similar results if we use other grades, such as fourth grade test scores.

reflection bias. If there are no pure class effects ($var(t_{cn} - t_n) = 0$) and peers do not affect earnings ($\beta = 0$),

$$plim\, \hat{b}^{LM} = \frac{\rho\theta}{\frac{1}{1-\frac{1}{I}} + \frac{2\theta+\theta^2}{1-\frac{1}{IC}}} \simeq \frac{\rho\theta}{(1+\theta)^2} \tag{24}$$

Using equation (24) for $\hat{b}_s^{LM}$ and the estimate of $\hat{\rho}_s$, we obtain an estimate of the reflection bias parameter $\frac{\theta}{(1+\theta)^2} = \hat{b}_s^{LM}/\hat{\rho}_s$. Combining this estimate and the estimate $\hat{\rho}_e$, we can then use equation (24) for $\hat{b}_e^{LM}$ to obtain an upper bound on the $\hat{b}_e^{LM}$ that could arise solely from reflection bias.

We implement the bound empirically by estimating the relevant parameters conditional on the vector of parent and student demographics, using regression specifications that parallel those used in column 3 of Table IV and column 2 of Table VIII. For eighth grade scores, we estimate $\hat{b}_s^{LM} = 0.057$ (SE $= 0.036$) and $\rho_s = 0.597$ (SE $= 0.016$), and hence

$$\frac{\theta}{(1+\theta)^2} = \frac{0.057}{0.597} = 0.0955.$$

For earnings, we estimate $\rho_e = \$90.04$ (SE $= \$8.65$) in Table IV. Hence, if the entire effect of class quality on earnings were due to reflection bias, we would obtain

$$\hat{b}_e^{LM} = \frac{\rho_e\theta}{(1+\theta)^2} = \$90.04 \cdot 0955 = \$8.60 \text{ (SE} = \$5.49)$$

where the standard error is computed using the delta method under the assumption that the estimates of $\hat{b}_s^{LM}$, ρ_s, and ρ_e are uncorrelated. This upper bound of \$8.60 due to reflection bias is only 17% of the estimate of $\hat{b}_e^{LM} = \$50.61$ (SE $= \$17.45$) in Table VIII. Note that the degree of reflection bias would be smaller in the presence of class quality effects ($\beta > 0$); hence, 17% is an upper bound on the degree of reflection bias in a linear-in-means model of peer effects.

Appendix C: Cost-Benefit Analysis

We make the following assumptions to calculate the benefits of the policies considered in the conclusion. First, following Krueger (1999), we assume a 3% annual discount rate and discount all earnings streams back to age 6, the point of the intervention. Second, we use the mean wage earnings of a random sample of the U.S. population in 2007 as a baseline earnings profile over the lifecycle. Third, because we can observe earnings impacts only up to age 27, we must make an

assumption about the impacts after that point. We assume that the percentage gain observed at age 27 remains constant over the lifecycle. This assumption may understate the total benefits because the earnings impacts appear to grow over time, for example as college graduates have steeper earnings profiles. Finally, our calculations ignore non-monetary returns to education such as reduced crime. They also ignore general equilibrium effects: increasing the education of the population at large would increase the supply of skilled labor and may depress wage rates for more educated individuals, reducing total social benefits. Under these assumptions, we calculate the present-value earnings gains for a classroom of 20 students from three interventions: improvements in classroom quality, reductions in class size, and improvements in teacher quality.

(1) Class Quality. The random-effects estimate reported in column 4 of Table VII implies that increasing class quality by one standard deviation of the distribution within schools raises earnings by $1,520 (9.6%) at age 27. Under the preceding assumptions, this translates into a lifetime earnings gain of approximately $39,100 for the average individual. This implies a present-value benefit of $782,000 for improving class quality by one within-school standard deviation.

(2) Class Size. We calculate the benefits of reducing class size by 33% in two ways. The first method uses the estimated earnings gain from being assigned to a small class reported in column 5 of Table V. The point estimate of $4 in Table V translates into a lifetime earnings gain from reducing class size by 33% for one year of $103 in present value per student, or $2,057 for a class that originally had twenty students. But this estimate is imprecise: the 95% confidence interval for the lifetime earnings gain of reducing class size by 33% for one year ranges from -$17,500 to $17,700 per child. Moreover, the results for other measures such as college attendance suggest that the earnings impact may be larger in the long run.

To obtain more precise estimates, we predict the gains from class size reduction using the estimated impact of classroom quality on scores and earnings. We estimate that a 1 percentile increase in class quality raises test scores by 0.66 percentiles and earnings by $50.6. This implies an earnings gain of $76.67 per percentile (or 13.1% per standard deviation) increase in test scores. We make the strong assumption that the ratio of earnings gains to test score gains is the same for changes in class size as it is for improvements in class quality more generally.[85] Under this

[85] This assumption clearly does not hold for all types of interventions. As an extreme example, raising test scores by cheating would be unlikely to yield an earnings gain of $77 per percentile improvement in test scores. The $77 per percentile measure should be viewed as a prior estimate of the expected gain when evaluating interventions such

assumption, smaller classes (which raised test scores by 4.8 percentiles) are predicted to raise earnings by $4.8 \times \$76.7 = \$368$ (2.3%) at age 27. This calculation implies a present value earnings gain from class size reduction of $9,460 per student and $189,000 for the classroom.

Calculations analogous to those in Krueger (1999) imply that the average cost per child of reducing class size by 33% for 2.14 years (the mean treatment duration for STAR students) is $9,355 in 2009 dollars.[86] Our second calculation suggests that the benefit of reducing class size might outweigh the costs. However, we must wait for more time to elapse before we can determine whether the predicted earnings gains based on the class quality estimates are in fact realized by those who attended smaller classes.

(3) Teachers. We calculate the benefits of improving teacher quality in two ways. The first method uses the estimated earnings gain of $57 from being assigned to a kindergarten teacher with one year of extra experience, reported in Figure IIIb. The standard deviation of teacher experience in our sample is 5.8 years. Hence, a one standard deviation increase in teacher experience raises earnings by $331 (2.1%) at age 27. This translates into a lifetime earnings gain of $8,500 in present value, or $170,000 for a class of twenty students.

The limitation of the preceding calculation is that it is based upon only one observable aspect of teacher quality. To incorporate other aspects of teacher quality, we again develop a prediction based on the impacts of class quality on scores and earnings. Rockoff (2004), Rivkin, Hanushek, and Kain (2005), and Kane and Staiger (2008) use datasets with repeated teacher observations to estimate that a one standard deviation increase in teacher quality raises test scores by approximately 0.2 standard deviations (5.4 percentiles). Under the strong assumption that the ratio of earnings gains to test score gains is the same for changes in teacher quality and class quality more broadly, this translates into an earnings gain of $5.4 \times \$76.7 = \$416$ (2.6%) at age 27. This implies a present-value earnings gain of $10,700 per student. A one standard deviation improvement in teacher quality in a single year generates earnings gains of $214,000 for a class of twenty students.

as class size or teacher quality for which precise estimates of earnings impacts are not yet available.

[86] This cost is obtained as follows. The annual cost of school for a child is $8,848 per year. Small classes had 15.1 students on average, while large classes had 22.56 students on average. The average small class treatment lasted 2.14 years. Hence, the cost per student of reducing class size is (22.56/15.1-1)*2.14*8848= $9,355.

References

Aaronson, Daniel, Lisa Barrow, and William Sander, "Teachers and Student Achievement in Chicago Public High Schools," *Journal of Labor Economics* 24:1 (2007), 95-135.

Almond, Douglas, and Janet Currie, "Human Capital Development Before Age Five," forthcoming, *Handbook of Labor Economics*, Volume 4 (2010).

American Community Survey, (http://www.census.gov, U.S. Census Bureau), 2006-2008 ACS 3-year data.

Angrist, Joshua D., Guido W. Imbens, and Alan B. Krueger. "Jackknife Instrumental Variables Estimation," *Journal of Applied Econometrics* 14:1 (1999), 57–67.

Angrist, Joshua D. and Alan B. Krueger, "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business and Economic Statistics*, American Statistical Association, 13:2 (1995), 225-235.

Angrist, Joshua D. and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press (2009).

Bacolod, Marigee P, "Do Alternative Opportunities Matter? The Role of Female Labor Markets in the Decline of Teacher Quality," *Review of Economics and Statistics,* 89:4 (2007), 737-751.

Chetty, Raj and John N. Friedman. "Does Local Tax Financing of Public Schools Perpetuate Inequality?" Forthcoming, *National Tax Association Proceedings* (2011).

Chetty, Raj, John N. Friedman, and Jonah Rockoff. "The Impact of Teacher Value Added on Student Outcomes in Adulthood" Harvard Univ. mimeo (2011).

Cilke, James "A Profile of Non-Filers," U.S. Department of the Treasury, Office of Tax Analysis Working Paper No. 78, July, 1998.

Corcoran, Sean P., William N. Evans, Robert M. Schwab, "Changing Labor-market Opportunities for Women and the Quality of Teachers, 1957-2000," *American Economic Review,* 94 (2004), 230-235.

Currie, Janet. "Inequality at Birth: Some Causes and Consequences." NBER Working Paper No. 16798, 2011.

Currie, Janet, and Duncan Thomas, "Early Test Scores, School Quality and SES: Longrun Effects of Wage and Employment Outcomes," *Worker Wellbeing in a Changing Labor Market*, 20 (2001), 103–132.

Dee, Thomas S., "Teachers, Race, and Student Achievement in a Randomized Experiment," *Review of Economics and Statistics,* 86 (2004), 195-210.

Dee, Thomas S., and Martin West, "The Non-Cognitive Returns to Class Size," *Educational Evaluation and Policy Analysis*, 33 (2011), 23-46.

Finn, Jeremy D., DeWayne Fulton, Jayne Zaharias, and Barbara A. Nye, "Carry-Over Effects of Small Classes," *Peabody Journal of Education,* 67 (1989) 75-84.

Finn, Jeremy D., Jayne Boyd-Zaharias, Reva M. Fish, and Susan B. Gerber, "Project STAR and Beyond: Database User's Guide," Lebanon: Heros, inc., 2007.

Guryan, Jonathan, Kory Kroft and Matthew J. Notowidigdo, "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments," *American Economic Journal: Applied Economics*, 1 (2009), 34-68.

Haider, Steven, and Gary Solon, "Life-cycle variation in the Association Between Current and Lifetime Earnings," *The American Economic Review,* 96 (2006), 1308-1320.

Hanushek, Eric A., "The Failure of Input-Based Schooling Policies." *Economic Journal* 113(1): F64-F98, 2003.

Hanushek, Eric A., "Economic Aspects of the Demand for Teacher Quality," prepared for the *Economics of Education Review,* 2010.

Heckman, James J., "Policies to Foster Human Capital," *Research in Economics*, 54:1 (2000), 3-56.

Heckman, James J., Jora Stixrud, and Sergio Urzua, "The Effects of Cognitive and Non-cognitive Abilities on Labor Market Outcomes and Social Behaviors." *Journal of Labor Economics* 24:3 (2006), 411-482.

Heckman, James J., Lena Malofeeva, Rodrigo Pinto, and Peter A. Savelyev, "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," unpublished manuscript, University of Chicago (2010).

Holland, Paul W. "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81 (1986), 945-960.

Hoxby, Caroline M. and Andrew Leigh, "Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States," *American Economic Review,* 94 (2004) 236-240.

Internal Revenue Service. *Document 6961: Calendar Year Projections of Information and Withholding Documents for the United States and IRS Campuses 2010-2018,* IRS Office of Research, Analysis, and Statistics, Washington, D.C, 2010.

Jacob, Brian A., Lars Lefgren and David Sims, "The Persistence of Teacher-Induced Learning Gains," Forthcoming, *Journal of Human Resources* (2011).

Jepsen, Christopher and Steven Rivkin, "Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size," Journal of Human Resources, 44:1 (2009) 223-250.

Kane, Thomas, and Douglas O. Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," NBER Working Paper No. 14607, 2008.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica,* 75 (2007), 83-119.

Krueger, Alan B, "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics,* 114 (1999), 497-532.

Krueger, Alan B., and Diane M. Whitmore, "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," *The Economic Journal,* 111 (2001), 1-28.

Lindqvist, Erik and Roine Vestman. "The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment." *American Economic Journal: Applied Economics*, 3 (2011), 101-28.

Manski, Charles, "Identification of Exogenous Social Effects: The Reflection Problem," *Review of Economic Studies,* 60 (1993), 531-542.

Muennig, Peter, Gretchen Johnson, Jeremy Finn, and Elizabeth Ty Wilde, The Effect of Small

Class Sizes on Mortality Through Age 29: Evidence From a Multi-Center Randomized Controlled Trial, unpublished mimeo, 2010.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges, "How Large are Teacher Effects?" *Educational Evaluation and Policy Analysis,* 26 (2004), 237-257.

Rivkin, Steven. G., Eric. A. Hanushek, and John F. Kain, "Teachers, Schools and Academic Achievement," *Econometrica,* 73 (2005), 417–458.

Rockoff, Jonah E., "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economics Review,* 94 (2004), 247-252.

Rockoff, Jonah E., and Douglas Staiger, "Searching for Effective Teachers with Imperfect Information," *Journal of Economic Perspectives,* 24 (2010), 97-117.

Sacerdote, Bruce, "Peer Effects with Random Assignment: Results for Dartmouth Roommates," *Quarterly Journal of Economics*, 116 (2001), 681-704.

Schanzenbach, Diane W., "What Have Researchers Learned From Project STAR?" *Brookings Papers on Education Policy,* (2006), 205-228.

Sims, David, "Crowding Peter to Educate Paul: Lessons From a Class Size Reduction Externality," *Economics of Education Review* 28:4 (2009), 465-473.

US Census Bureau. "School Enrollment–Social and Economic Characteristics of Students: October 2008, Detailed Tables," Washington, D.C., 2010.

(http://www.census.gov/population/www/socdemo/school.html).

Word, Elizabeth., John. Johnston, Helen. P. Bain, B. Dewayne Fulton, Charles M. Achilles, Martha N. Lintz, John Folger, and Carolyn Breda, "The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985–1990," Tennessee State Department of Education, 1990.

III Why Don't People Trust Experts?[87]

III.A Introduction

Your doctor probably knows more about what medical treatments you need than you do. She also has a financial stake in the decision of which treatment to provide. This is called the "credence good" problem. Sellers of credence goods are known as "experts."

Credence goods are closely related to the "lemons" analyzed in Akerlof (1970). Whereas Akerlof analyzed the decision of whether to trade a single good, credence goods generalize this problem to the selection of one good to trade out of multiple goods on a menu. In the one-good model, the seller has private information about the buyer's price-conditional surplus for the single good; in the credence good model, the buyer has private information about the buyer's price-conditional ranking over all goods on the menu.

This generalization creates a familiar problem. Whereas the market for lemons only suffers from foregone trade, the market for credence goods suffers from (socially) sub-optimal choices of goods from the expert's menu, due to biased recommendations of the expert. This is known as mistreatment. Mistreatment of patients by doctors, for example, is thought to be a major source of large geographic cost variation in Medicare (Skinner 2009).

A recent review article by Dulleck and Kerschbamer (2006, henceforth DK) clarifies the diverse body of existing results on credence goods. They show, surprisingly, that the literature has so far failed to provide any simple, robust explanation for widespread mistreatment. They conclude that "...a comparison of common experience with the results of this paper suggests that something is missing from existing models of credence goods."

In this paper, I provide a simple rationale for mistreatment: consumers do not perfectly observe expert cost functions. This point has a long history in the literature on regulation of monopolists (Baron and Myerson 1982)[88] but has not been fully recognized in the credence good literature.

[87]I would like to thank Oliver Hart for numerous insightful suggestions, as well as overall guidance and support. I also thank Philippe Aghion, Raj Chetty, Tyler Cowen, David Cutler, Melissa Eccleston, Drew Fudenberg, Joshua Gottlieb, Andrei Shleifer, and Danny Yagan for helpful comments. Funding from a National Science Foundation Graduate Research Fellowship and Harvard University are gratefully acknowledged.

[88]Baron-Myerson (1982) assume that sellers only possess private information about supply, not demand. This is realistic when regulating the production of one large product (e.g., a bridge) to a community, but not when regulating the production of thousands of small products (e.g., medical treatments) tailored to individual consumer circumstances.

Under the more traditional and realistic assumption that consumers do not observe seller costs, mistreatment in both directions becomes a deep problem, robust to endogenous pricing, competition, and endogenous technology-adoption by experts.

DK rely heavily on the assumption that expert cost functions are common knowledge in their analysis. They recognize this limitation, but do not explore alternate assumptions, and do not recognize how radically their common-knowledge assumption departs from the adverse selection and regulation literatures on asymmetric information. In reality, expert cost functions are far from common knowledge. An expert's cost function may depend on local prices of capital and labor, the form of organization, production capacity, and her own skills and preferences. Consumers typically have at best a rough estimate of expert cost functions. A consumer with heart disease does not know how much it costs *any* heart surgeon to provide an echocardiogram, stents, or rotoblation– much less the consumer's particular surgeon, at a particular point in time. Even sophisticated buyers of expert services, such as HMO's, observe only a small fraction of the actual variation in cost functions across individual suppliers. Moreover, cost functions can vary endogenously. Given a set of market prices, or a reimbursement rule fixing prices at some local average cost as in Medicare, experts may partly specialize in particular treatments to maximize the profits from mistreatment.

My analysis ignores what I refer to as three important "non-price constraints" on experts: professional ethics, reputation, and second-opinions. All of these constraints will tend to reduce problems of asymmetric information, including the adverse selection that arises in Akerlof (1970), the information rents that arise in Baron-Myerson (1982), and the mistreatment that arises in credence good markets. However, continuing the tradition of these earlier authors, I ignore these constraints because they seem likely to complicate the analysis for gains of minor additional insight[89]. Below I also discuss many reasons why these non-price constraints are unlikely to solve the mistreatment problem completely.

Understanding why mistreatment arises in credence good markets is important because a large and growing share of expenditures involve credence goods. Leading examples of credence good sellers include physicians, money managers, lawyers, mortgage brokers, real estate agents, home and auto repair experts, and funeral parlors. In addition, many public goods exhibit credence. Credence

[89] However, see Ely and Valimaki (2003) for an example in which the intuition that reputation alleviates information problems turns out to be false.

arises when politicians possess inside information about the social costs or benefits of potential public expenditures; politicians then act as experts "selling" goods to voter-taxpayers. Similarly, CEOs of corporations act as experts, and voter-shareholders pay for spending with foregone dividends.

Understanding mistreatment in theory is also important because empirical evidence suggests it may be an economically large problem. A large literature in health economics tests for overtreatment, referred to as "physician-induced demand" (PID). The literature finds evidence of mistreatment but suffers from identification problems (McGuire 2000). More recent studies of variation in Medicare spending (Fisher *et al* 2003a,b, Skinner 2009a,b, Skinner *et al* 2006, Baicker and Chandra 2004, Sirovich 2008, Gruber and Owings 1996, Gawande 2009) suggest that mistreatment may be economically significant relative to medical spending growth[90].

The leading explanation for geographic variation in medical spending is geographic variation in capacity. This explanation points to unobserved variation in expert cost functions[91] as the key source of inefficiency. While previous authors have explored the role of idle capacity in credence good markets (Emons 1997 and 2001, Richardson 1999), no one has analyzed the more general problem of unobserved cost functions, shown how robustly they push the market toward mistreatment, or placed this problem in its proper context in the asymmetric information literature.

Akerlof (1970) pointed out that governments can "solve" the lemons problem with product "pooling," or a mandate that all sellers sell their goods at identical prices. Similarly, governments can "solve" the credence good problem with input pooling, or a mandate that all sellers buy identical inputs at identical prices. In theory, this would render all experts' cost functions common knowledge. In practice, just as it is costly and difficult to force sellers of different goods to adopt identical sale prices (e.g., to force healthy and sick people to "sell" their risk portfolio to health-insurers for the same price), it is costly and difficult to force sellers to adopt identical input mixes at identical input prices (e.g., to force better and worse doctors to buy the same "effort" input quan-

[90] Studies have documented overtreatment in other markets, as well. Jessica Mitford (1998) informally documents overtreatment by funeral good sellers. Studies of auto mechanics by the Department of Transportation (New York Times 1979, cited in Wolinsky 1993) and of optometrists by the Federal Trade Commission (FTC 1980, cited in Wolinsky 1993) claim to find substantial overtreatment. All of these studies can be interpreted in the framework of unobserved costs presented here.

[91] It is useful to distinguish (1) PID, which is mistreatment as defined in this paper, from (2) excessive treatment due to fixed prices or monopoly quantity-setting (McGuire 2000). In (1), supply equals demand. In (2), supply does not equal demand. Presumably, the two cases can be distinguished empirically by observing correlation between consumer satisfaction and treatment rates. The lack of correlation observed in the existing literature (Fisher et al 2003.b) suggests mistreatment, not expert-imposed quantity-setting against the will of consumers.

tities at the same implicit prices). Hiring experts on salary and purchasing inputs on their behalf, as do some hospitals and public health systems, can be viewed as attempts to move toward this pooling strategy. Input-pooling cannot fully equalize input prices for time, effort, and ability, but might help if non-price constraints function better when experts have less to gain from dishonesty. This loosely describes some of the more efficient healthcare producers in Gawande (2009).

The paper proceeds as follows. I first provide an example that captures the main story of the paper. I then present a model of credence goods with unobserved cost functions. I solve for separating and pooling equilibria under assumptions of monopoly and exogeneity of cost function type. I then extend the model to a competitive bidding process. I then endogenize the expert's choice of cost function, under both monopoly and competition. I then discuss the results and conclude.

III.B Note on Prior Literature

The results presented here fit into DK's framework in a very clear way. DK show in their Proposition 1 that under three simple assumptions, endogenous pricing by experts solves the credence good problem. These assumptions are (i) consumers are all the same (assumption H for "Homogeneity"), (ii) getting second-opinions is not possible (assumption C for "Commitment"), and (iii) either consumers can verify what good is provided (assumption V for "Verifiability") or sellers are liable to unlimited punishments for selling a good of inefficiently low quality (assumption L for "Liability"). Moreover, efficiency is maintained without H if there is competition, and efficiency is maintained without C if prices are not restricted. If assumptions V and L are both relaxed then all experts either "overcharge" consumers, i.e. recommend expensive goods and secretly provide inexpensive goods, or there is no trade.

Unobserved cost functions escape the efficiency result of DK's Proposition 1 through a partial relaxation of assumption V. Suppose a consumer with a hand problem may need expensive surgery, or may only need inexpensive cortizone shots. It is implausible in this example, and most other medical examples, to suppose that the consumer cannot verify which treatment she has actually received. But if we view the same treatment provided at different costs as different treatments, then partial relaxation of V becomes realistic. The consumer can observe that she has received surgery,

but cannot tell how much it cost the expert to provide it. Finally, in keeping with DK's Proposition 1, I must also relax assumption L. A partial relaxation of L is sufficient for my results, and realistic; experts who undertreat consumers face only finite and perhaps small expected punishments for a number of reasons. However, to simplify the algebra below, I fully relax L.

DK suggest a different explanation for the failure of existing models to explain consumer fears of mistreatment, i.e., the sense that "something is missing from existing models of credence goods." They suggest that consumers may be unwilling to pay equal markups on treatments with very different prices. To use their example, if an auto mechanic has a choice between recommending a new fuse that costs the expert $20 or a new engine that costs the expert $3000, and the expert stands to gain $500 in profit from installing a new engine because he will use idle capacity that would otherwise generate zero profit, then equal markups require a $500 markup on a $20 fuse. DK suggest that a $520 fuse would outrage consumers. But this is not obvious. Such consumers would get cheaper treatments when they actually did need them, in effect cross-subsidizing themselves from good states to bad states over time. In the limit, this would amount to an annual subscription fee, with payment for all needed services at cost, i.e., a standard maintenance or insurance plan. Unobserved cost functions appear more likely to be the missing "something" than social norms against collecting insurance fees for risks that have not eventuated.

For the sake of clarity, I should note that the credence good model here does not fully generalize the lemons model to a multiple-good setting. In the lemons model, it costs the expert seller more to provide the version of the single good with higher consumer valuation (the non-lemon). In the credence good model here, it does not. This is because the "quality" of the good is here defined by the consumer's state, not any inherent characteristic of the good. This assumption would eliminate adverse selection in the one-good model, because equilibrium prices under pooling give sellers of high-quality goods no greater incentive to withdraw from the market than sellers of low-quality goods. The fully-general choose-one-out-of-many lemons problem is more complicated than we need for credence goods, and would involve both adverse selection and mistreatment. Credence goods still involve private information about both supply (a seller's cost function) and demand (the consumer's valuations of all goods on the menu). But the credence good structure reduces the amount of information there is to know about supply by holding a given seller's costs constant for a particular good, even as different consumer rankings of that good can vary. The credence good

model therefore eliminates variation that underlies adverse selection in order to focus on the *new* problem that arises under multiple goods: mistreatment.

Finally, the approach taken here may appear similar to that taken by Fong (2005). Fong endows the expert with private information on which consumers are more expensive to treat. This results in experts *overcharging* some consumers, i.e., pretending to provide expensive treatments while only providing inexpensive treatments. While endowing sellers with private information on which consumers are more expensive to treat is similar to the assumption explored here (that sellers possess private information on their own cost functions), Fong's argument is less appealing for several reasons.

First, the inefficiency Fong derives is overcharging, not mistreatment. This is not a realistic form of inefficiency in many important cases. For example, virtually all concerns about the health care market involve mistreatment, not mischarging. Second, descriptions of the consumer heterogeneity that generate Fong's overcharging result are not as simple and appealing as the more traditional story of unobserved cost functions. For example, he states, "...a crack on the muffler of a car will not bother much a car owner with bad hearing. However, if he is a Hi-fi enthusiast and has installed an expensive car stereo, he will suffer much more." Third, Fong derives inefficiencies in mixed strategy equilibria, which are harder to interpret than the pure-strategy equilibria derived here. Fourth, he makes several other complicated assumptions. He assumes that experts cannot price discriminate across heterogeneous consumers despite being able to recognize them well enough to target them with overcharging; that consumers cannot hide their own types from experts or send other types on their behalf; that it is not worthwhile for consumers to fix a small problem with a large treatment; and that assumptions V, H, and C all simultaneously break down in very specific ways. In contrast, the DK-based model presented here is simple, easy to explain with realistic examples, and yields robust predictions of mistreatment (not mis*charging*) as pure strategies.

III.C An Example

As above, consider a risk-neutral consumer with a hand problem. The consumer is willing to pay 15 for healthy hands. But he doesn't know if he has carpal tunnel syndrom, which requires major surgery, or minor inflammation, which only requires cheap cortizone injections. She knows each problem is equally likely.

There are two types of doctors, S and B. S doctors perform surgery for carpal tunnel syndrome more easily than B doctors; surgery costs S doctors 8, and B doctors 10. Both doctors can provide cortizone shots at the cost of 4. The "cost wedge" of a doctor is the difference between her costs of providing the more and less expensive treatments. Thus S stands for "smaller cost wedge" and B stands for "bigger cost wedge."

If patients know what kind of doctor they're dealing with, they can calculate the markups implied by prices and costs. Thus an S doctor charging 13 for cortizone shots and 17 for surgery would earn the same profit on both treatments. Consumers could recognize this and trust the doctor to provide the right treatment, since the doctor's profit does not depend on her diagnosis. The same holds for a B doctor charging 12 for cortizone and 18 for surgery. Since patients are willing to pay more for an honest doctor, and a monopolistic doctor gets to extract the patient's full WTP, this is the doctor's optimal pricing scheme when consumers observe cost functions.

But now, more realistically, suppose patients don't know what kind of doctor they're dealing with. Patients only know that type S and type B doctors are equally common.

Suppose each type of doctor adopts the same equal-markup strategy as before. The expected profit for S will be 9, and the expected profit for B will be 8, reflecting S's greater skill at performing surgery. Under this candidate equilibrium, consumers will infer the type of doctor they're dealing with from posted prices: the price vector $(13, 17)$ will induce consumers to believe a doctor is type S, while a price vector $(12, 18)$ will induce consumers to believe a doctor is type B.

Is this incentive-compatible? To find out, we must calculate the profit that S doctors can earn from imitating B doctors, and vice versa. If one type of doctor imitates another, his markups across treatments will vary, and he will mistreat consumers: S firms charging honest-B prices will always perform surgery, and B firms charging honest-S prices will always provide cortizone shots. Unfortunately, this yields profits of 10 for S and 9 for B. Thus both types of doctors would deviate from the honest equilibrium by imitating each other's prices and mistreating consumers. Below, I show that this is a general feature of the credence good situation when consumers do not observe cost functions.

What equilibrium will emerge in this example, if not the honest one? Two salient candidates for equilibria are an overtreatment pooling equilibrium at price $(12, 18)$ in which S doctors always perform surgery but B doctors are honest, and an undertreatment pooling equilibrium at price

$(13, 17)$ in which B doctors always provide cortizone shots but S doctors are honest. However, the consumer now anticipates mistreatment, lowering her WTP for expert services. Therefore, the first step is to reduce these prices while preserving equal markups for the relevant types. As I show below, the proper reductions depend nonlinearly on the consumer's WTP to solve her problem, the probability of having the big problem, the probability that the expert has type B cost function, and the cost wedges. In this example, the prices that emerge when S pools with honest-B to yield partial overtreatment are $(10.5, 16.5)$ and the prices that emerge when B pools with honest-S to yield partial undertreatment are $(10.25, 14.25)$.

Consider the pooling overtreatment equilibrium at prices $(10.5, 16.5)$, at which B earns honest profits 6.5 and S earns overtreatment profits 8.5. (To preview later results for the case when experts can choose their cost functions, note that it is not a coincidence that the dishonest doctor earns higher profit than the honest doctor.) To sustain this, each doctor must consider and reject many alternative price vectors. In calculating profits at price vectors off the equilibrium path, doctors must impute beliefs to consumers. For example, in the overtreatment equilibrium, B must not be tempted to announce the off-equilibrium price vector $(13, 17)$, the separating-equilibrium honest price vector for S.

If consumers infer that such an announcement must be made by an S doctor, then B earns a profit of 9 and hence deviates: he gets to undertreat consumers, and consumers have a high WTP because they think they're being treated honestly (by an S doctor at honest-for-S prices). So clearly consumers cannot react to such a surprise announcement with these beliefs if we are to sustain the equilibrium.

If consumers infer that the $(13, 17)$ announcement must be made by a B doctor, thereby anticipating undertreatment, they will reject the contract and both S and B will earn zero profits. The more useful deviation to test when consumers believe the deviator is a B doctor, then, is the maximum price vector B can charge that still induces trade in this family of deviations, where the family of deviations is defined as price vectors for which an S doctor would earn equal markups. This is just one family of deviations. In the proofs, I partition the price space into distinct families of deviations, and examine the most profitable trade-inducing deviations given consumer beliefs for each type of doctor in each family of deviations. Here, in the "S prices honestly" region of the price space, when consumer's perceive deviators as B doctors, that maximum-profit price vector

is $(7.5, 11.5)$, leaving both S and B doctors with a profit of 3.5. Therefore this belief supports the equilibrium with respect to this family of deviations in this example. Further calculations show that there do exist beliefs at all off-equilibrium price vectors that support the overtreatment equilibrium in this example. In the proofs, general patterns emerge that show many of these calculations reduce to a single restriction on the parameter space with intuitive implications.

The inefficiency created by the overtreatment equilibrium is the expected excess expenditure on the big treatment, which occurs when the doctor has cost type S and the consumer has the small problem. In this example, the inefficiency is $.5 * .5 * (8 - 4) = 1$. This means a government maximizing total output should be willing to spend up to one unit of resources to eliminate unobserved cost variation.

Are the beliefs discussed so far consistent with the "Intuitive Criterion" of Cho and Kreps (1987)? The Intuitive Criterion requires that beliefs at a deviation place no weight on doctors for whom that deviation is weakly dominated over all beliefs by the equilibrium. Clearly, B does prefer $(13, 17)$ for the belief that S made the deviation (profit of 9, as shown above) so that belief is consistent with the Intuitive Criterion. At $(7.5, 11.5)$, on the other hand, beliefs do not affect profits–the consumer is willing to trade at all beliefs, and both doctors earn profits below the equilibrium, so the Intuitive Criterion places no restrictions on beliefs here. It turns out that the Intuitive Criterion does not affect the analysis. An even stricter equilibrium concept called Neologism Proofness discussed by Farell (1993) renders the coexistence of over- and undertreatment equilibria impossible, and in some cases leaves no pure-strategy equilibria at all. This is consistent with Farell's finding that NP equilibria need not exist.

Can the pooling undertreatment equilibrium also exist in this example? As before, the $(13, 17)$ prices that work under separation must be adjusted downward to reflect the consumer's lower WTP under pooling. The adjusted prices are $(10.25, 14.25)$. Similar calculations as above show that in this example the undertreatment equilibrium cannot be supported. The reason is that S doctors will deviate from a $(10.25, 14.25)$ price vector where S earns profits of 6.25 to, among other prices, a $(0, 15)$ price vector where S earns profits of 7 at all consumer beliefs (at these prices, profits do not depend on beliefs since consumers know that both types will overtreat). So in this example the only pure strategy equilibrium is the overtreatment equilibrium found above.

In other examples, both overtreatment and undertreatment equilibria can arise without addi-

tional refinements on beliefs such as Neologism-Proofness. While these refinements are interesting, they complicate the model and the comparative statics without adding subsantially to the argument, and I therefore ignore them in what follows.

III.D A Model of Hidden Cost Functions

III.D.1 Setup

The model extends the framework developed in DK (2006). To ease exposition, continue the above example and imagine a consumer is willing to pay v for healthy hands, but may require expensive surgery for carpal tunnel syndrome (the high treatment) at cost c_H or cheap cortizone shots (the low treatment) at cost c_L. The consumer needs surgery with probability h.

There are two types of doctors, S and B, defined as above by their cost vector $c_i = (c_{Li}, c_{Hi})$ with $i \in \{S, B\}$ and cost wedge $\Delta c_i = c_{Hi} - c_{Li}$. Define Δc_S and Δc_B as the smaller and larger cost wedges such that $\Delta c_S \leq \Delta c_B$. Define c_S and c_B as the analogous cost vectors and q_S, q_B as the analogous price vectors. Thus q_{HS} refers to the price charged for surgery by S doctors[92], and c_{LB} refers to the cost of providing L for a B doctor.

I assume that cost functions vary in a particular way: high and low costs are negatively correlated. This means that $c_{LB} \leq c_{LS}$ and $c_{HS} \leq c_{HB}$. This assumption describes a world with specialization by doctors in one treatment or the other, and rules out the world in which some doctors are simply better at everything than other doctors. This assumption is often plausible and is required to make the model interesting.

Denote the consumer's belief at an information set (price vector q) as $\mu \equiv \Pr(B|q)$ (omitting the q argument for simplicity), the perceived likelihood that a B doctor is offering the observed prices. Define the share of type B doctors in the market as $p = \Pr(c = c_B)$, or, in words, the probability that a given firm, drawn from a two-type cost distribution, has the cost-type with the bigger cost wedge.

The market has $n \geq 1$ risk-neutral experts simultaneously choosing prices, and measure 1 of

[92] It is important to keep the unobservability of cost functions in mind when interpreting the model. The imagined experts must be sufficiently similar in apparent realms of expertise that normal consumers cannot distinguish them. The distinction here, for example, should not be thought of as that between primary care physicians (PCPs) and surgeons. Consumers know that a surgeon has a cost advantage in providing surgery relative to a PCP. This is why the numerical example above differentiates cost functions by the more hidden metric of surgeon skill. As discussed in the text, subtle but substantial cost variation of this sort may arise for many reasons.

risk-neutral consumers.

I now restate the key definitions and assumptions employed in the analysis of DK (2006), already discussed above.

Define "overcharging" as supplying the low treatment but charging for the high treatment.

Define "verifiability" as observability of treatment type by consumers. I assume verifiability in the sense that consumers recognize the type of treatment they receive, but not in the sense that consumers can distinguish identical treatments by their cost to the expert. This assumption rules out overcharging: our hand doctor can't charge a consumer for surgery but then just inject cortizone instead.

Define "undertreatment" as providing the low treatment to a consumer with the big problem, and "overtreatment" as providing the high treatment to a consumer with the small problem. A doctor who likes to perform surgery tends toward overtreatment, while a doctor who considers surgery a hassle tends toward undertreatment.

Define "liability" as the restriction that undertreatment guarantees an infinite punishment of the expert. Liability ensures that firms will not undertreat. Below I assume there is no liability–the doctor who always injects cortizone, even when patients need surgery, never gets punished. The results are qualitatively similar under sufficiently limited liability, though they are not robust to unlimited liability[93].

Next, define "commitment" as the restriction that diagnosis by a firm entails the implied treatment by that firm. I assume commitment and hence ignore the constraints implied by second opinions.

I further assume that it is socially efficient for consumers to solve their problems, i.e., $v > c_H \geq c_H - c_L = \Delta c$ for all cost functions.

The order of play in the expert-consumer game is as follows. Nature chooses expert cost functions and consumer problem types. Experts learn their cost functions and choose prices q.

[93] Why does unlimited liability restore efficiency? In this case, the only mistreatment we need to consider is overtreatment. The firm tempted to overtreat has lower costs of providing expensive treatment, and attempts to pool on the higher-cost firm's pricing for the expensive treatment. But the higher-cost-for-expensive-treatment firm can simply decrease its price on the more expensive treatment and increase its price on the inexpensive treatment one-for-one until it is no longer profitable for the other firm to pool on pricing and overtreat. Full liability makes this possible by removing the temptation to undertreat as the markup on the inexpensive good gets arbitrarily large. This is why partial liability is sufficient for my results–it does not permit the higher-cost-for-expensive-treatment firm to shift enough profit into the less expensive good to eliminate the temptation of the lower-cost-for-expensive-treatment firm to copy the other firm's price and overtreat.

Consumers observe prices and choose whether or not to visit an expert. If the consumer does visit an expert, the expert diagnoses and treats the consumer. The game tree is depicted in Figure 3.1.

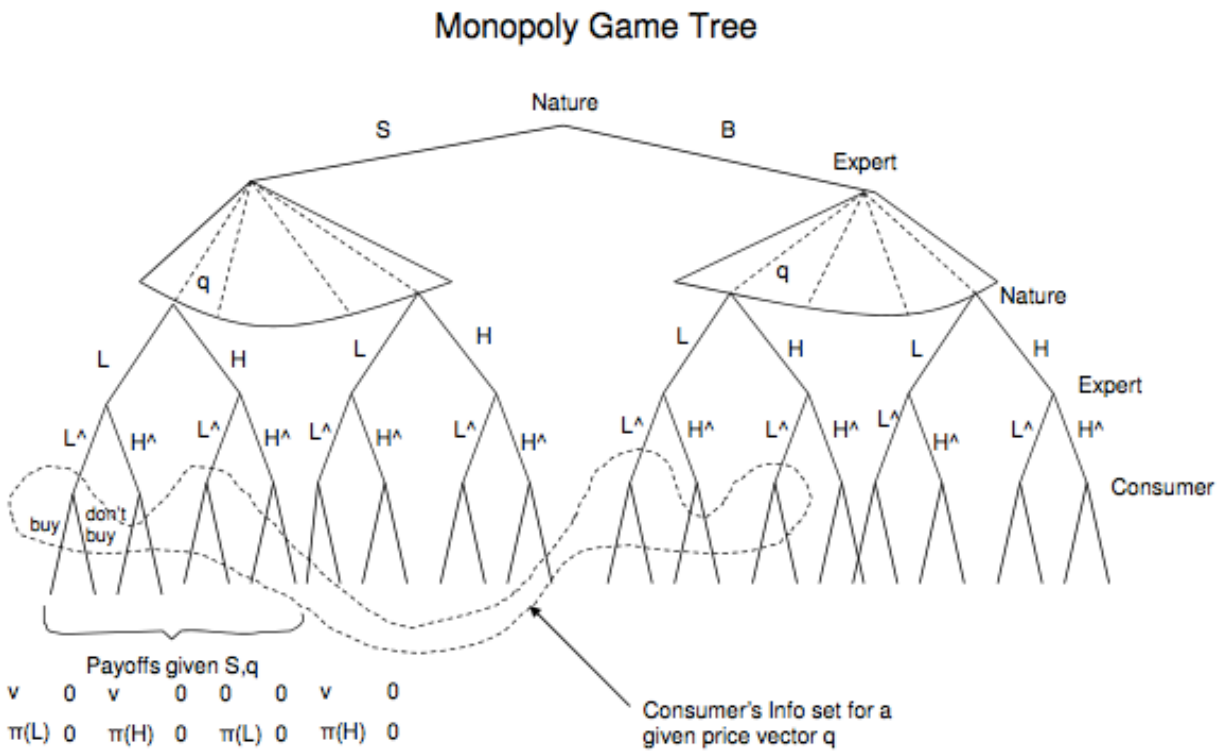

Monopoly Game Tree
Nature
S
B
Expert
q
q
Nature
L
H
L
H
L
H
L
H
Expert
L^
H^
L^
H^
L^
H^
L^
H^
L^
H^
L^
H^
L^
H^
L^
H^
Consumer
buy
don't buy
Payoffs given S,q
v 0 v 0 0 0 v 0
π(L) 0 π(H) 0 π(L) 0 π(H) 0
Consumer's Info set for a given price vector q

Figure 3.1

Much of the analysis that follows refers to the price space (q_H, q_L), depicted in Figure 3.2. Prices yield markups over costs and define the expert's incentive-compatibility (IC) line, $q_L = q_H - \Delta c$. Any point southeast of the IC line commits the expert to overtreatment. Any point northwest of this line commits the expert to undertreatment. Note that the price space is partitioned into five regions labeled $R1$-$R5$. Both types of experts will undertreat in $R1$ and overtreat in $R5$. Type S will behave honestly in $R2$ and overtreat in $R3$ and $R4$. Type B will behave honestly in $R4$ and undertreat in $R2$ and $R3$. Let $\pi_S(R2|\mu = 0)$ refer to the maximum profit function of type S for prices constrained to lie in $R2$, given that consumers believe that only type S will price in $R2$. Thus $\pi_S(R2|\mu = p)$ is the S expert's maximum profits in $R2$ when consumers believe that both expert types will choose the same price in $R2$. Experts can separate by pricing in different regions, or by pricing at different price vectors in the same region.

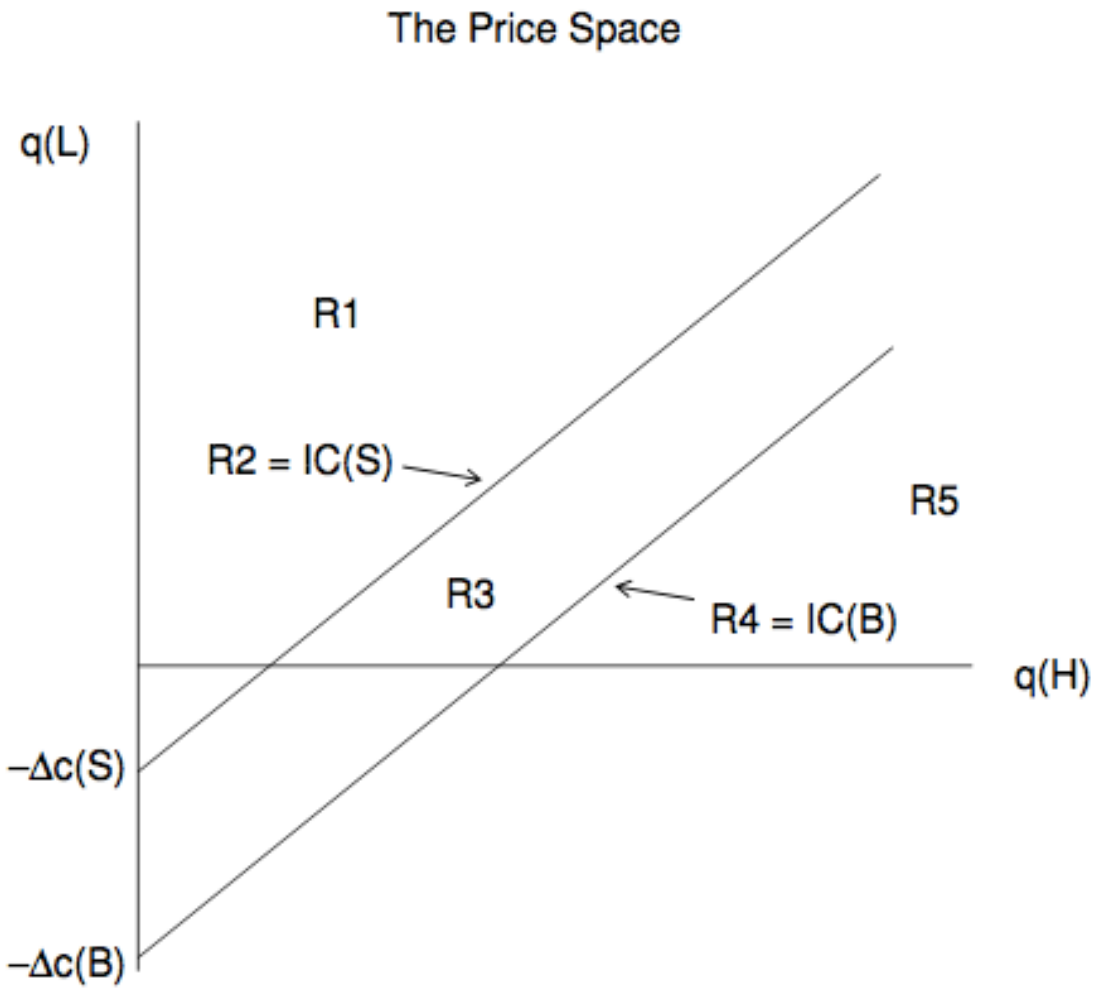


Figure 3.2

III.D.2 Observed Costs: Monopoly Case

I begin with the monopoly case in which the expert offers a take-it-or-leave it contract to the consumer. I call the "first best" equilibrium that which emerges when the consumer knows the expert's cost type. This separating equilibrium replicates the results from DK. The expert's strategy is a mapping from costs c into chosen prices q.

The first-best outcome is a fully separating equilibrium with honest treatment. Under separation, each firm solves its problem as if there were a single cost function. Hence we can apply Lemma 1 from DK separately for each type of firm. By Lemma 1, optimal prices satisfy $q_L + h\left(q_H - q_L\right) = v$ and $q_H - c_H = q_L - c_L$.

Solving for optimal prices as in DK, we get for $j \in \{S, B\}$:

$$q_{Lj} = v - h\Delta c_j \tag{25}$$

$$q_{Hj} = v + (1-h)\,\Delta c_j. \tag{26}$$

Profits are:

$$\pi_j = v - hc_{Hj} - (1-h)\,c_{Lj}. \tag{27}$$

Under separation, it is optimal for each type of expert to commit to honest treatment of consumers. Honesty maximizes consumers' willingness to pay (WTP) for expert services, and experts can set prices to extract full WTP.

III.D.3 Unobserved Costs: Monopoly Case

I refer to equilibria in which the expert has private information about her cost type as "second-best" equilibria. The following two propositions characterize all pure-strategy PBEs of the expert-consumer second-best monopoly game. They show that perfect honesty is not possible, but partial honesty is guaranteed. Note that profits for all regions and all relevant beliefs along the consumer's PC are listed in the appendix.

Proposition 1 *There is no pure strategy honest separating equilibrium in the expert-consumer second-best monopoly game.*

Proof. Suppose S and B both treat honestly. Then S maximizes profits in $R2$ given $\mu = 0$ and B maximizes profits in $R4$ given $\mu = 1$. First, suppose that S and B price on the consumer's PC. Then given these beliefs, B will prefer to pool with S in $R2$ iff

$$\pi_B(R4|\mu = 1) < \pi_B(R2|\mu = 0) \tag{28}$$

$$\iff v - hc_{HB} - (1-h)c_{LB} < v - h\Delta c_S - c_{LB} \tag{29}$$

$$\iff \Delta c_S < \Delta c_B, \tag{30}$$

which is true by assumption. Similarly, S will prefer to pool with B in $R4$. Now suppose that either S or B prices strictly inside the consumer's PC, leaving the consumer with positive surplus. For S not to pool with B, the difference $\Delta q_H \equiv q_{HB} - q_{HS}$ must decrease. For B not to pool with S, the difference $\Delta q_L \equiv q_{LS} - q_{LB}$ must decrease. But since we have the relation $\Delta q_H + \Delta q_L = \Delta c_B - \Delta c_S =$ constant, this is not possible. QED. ■

Proposition 2 *There exist only two pure-strategy PBE in the expert-consumer second-best monopoly game. Suppose*

$$v > \Delta c_B \left[1 + \frac{(1-h)(1-p)}{h}\right]. \tag{31}$$

Then there exists an "overtreatment pooling" PBE at prices $q_L = v - [1 - p(1-h)]\Delta c_B$ *and* $q_H = v + p(1-h)\Delta c_B$*, and profits* $\pi_S = v + p(1-h)\Delta c_B - c_{HS}$ *and* $\pi_B = v - (1-h)(1-p)\Delta c_B - hc_{HB} - (1-h)c_{LB}$*, in which* S *treats all consumers for the big problem, and* B *treats consumers honestly. Now suppose*

$$v < \Delta c_S \left[1 + \frac{1-h}{ph}\right]. \tag{32}$$

Then there exists an "undertreatment pooling" PBE at prices $q_L = (1-ph)v - h(1-p)\Delta c_S$ *and* $q_H = (1-ph)v + [1 - h(1-p)]\Delta c_S$*, and profits* $\pi_S = (1-ph)v + ph\Delta c_S - hc_{HS} - (1-h)c_{LS}$ *and* $\pi_B = (1-ph)v - h(1-p)\Delta c_S - c_{LB}$ *in which* B *treats all consumers for the small problem, and* S *treats consumers honestly.*

Proof. See Appendix. ■

Proposition 2 has some interesting implications. These implications should be viewed as merely suggestive due to the simplicity of the model.

1. All equilibria involve some honest and some dishonest experts. We never have all crooks or all saints.

2. The dishonest firm earns higher profits. Therefore profits are positively correlated with mistreatment.

3. Only firms specializing in the costlier treatment ever overtreat, and only firms specializing in the cheaper treatment ever undertreat.

4. Market prices depend on the cost wedge of honest firms, but not the cost wedge of dishonest firms. In particular, a larger cost wedge of honest firms lowers q_L and raises q_H for both honest and dishonest firms.

5. Overtreatment and undertreatment equilibria can both exist in parts of the parameter space (multiple equilibria), but only if firms are not "too different," i.e., only if $\frac{\Delta c_S}{\Delta c_B}$ is close enough to 1, and only if consumers are not too likely to have the big problem, i.e., only if h is small enough.

6. We get clean comparative statics on the direction of mistreatment with respect to all exogenous parameters. Define indicator variable I_O to equal 1 if a market can support an overtreatment equilibrium and otherwise 0, and I_U to equal 1 if a market can support an undertreatment equilibrium and otherwise 0. Define the parameter space Ω. The five exogenous parameters in Ω are $v, h, p, \Delta c_S, \Delta c_B$. As shown, the implications of the model for the likelihood of overtreatment versus undertreatment at this stage are:

$$\frac{\partial I_O}{\partial v} \geq 0, \frac{\partial I_O}{\partial p} \geq 0, \frac{\partial I_O}{\partial h} \geq 0, \frac{\partial I_O}{\partial \Delta c_S} = 0, \frac{\partial I_O}{\partial \Delta c_B} \leq 0 \tag{33}$$

$$\frac{\partial I_U}{\partial v} \leq 0, \frac{\partial I_U}{\partial p} \leq 0, \frac{\partial I_O}{\partial h} \leq 0 \ , \frac{\partial I_U}{\partial \Delta c_S} \geq 0, \frac{\partial I_U}{\partial \Delta c_B} = 0. \tag{34}$$

Overtreatment is more likely in a market where experts fix more important problems (bigger v), and when big problems more commonly befall consumers (bigger h). If consumers place a high value on solving their problems, and these problems tend to require major treatment, pricing plans that induce overtreatment will allow firms to extract much more surplus and therefore tend to prevail. Medicine comes to mind as a good candidate for this kind of market, at least in terms of

large v. The model suggests that overtreatment will be more common for expensive treatments fixing more common problems, such as heart disease, than less common problems, such as Lou Gehrig's Disease. It also suggests that overtreatment of a particular problem will be more common in populations with higher risks of needing expensive treatment, such as bed bug extermination in New York City, and Alzheimer's Disease medication among older people.

Overtreatment is also more likely when there is a greater chance that the firm is better at the cheap treatment. This is because if all firms are better at the cheap treatment, the overtreatment equilibrium converges to an honest equilibrium – consumers only get ripped off by a few bad apples who are "too good" at the expensive treatment. This may seem reminiscent of specialty clinics that offer big treatments like heart surgery or Lasik eye surgery at discount prices, and allegedly tend to overtreat consumers. However, this application of the model is incorrect if such firms are so different from nonspecialists that consumers perceive this difference.

Overtreatment is also more likely when Δc_B is small, but oddly enough does not depend on Δc_S. Symmetrically, undertreatment is more likely when Δc_S is big, but does not depend on Δc_B. The idea is that in each equilibrium, the binding temptation is for the honest firm to deviate to a pure-mistreatment pricing vector, in the direction in which he specializes. When Δc_B is small, the B firm has a smaller cost advantage in the cheaper treatment and so is less tempted to deviate, making overtreatment easier to sustain. This temptation is not affected by Δc_S because Δc_S has no effect on the profits that would be gained at this alternative pure-undertreatment price. These results are harder to map into real phenomena.

Which is worse, undertreatment or overtreatment? The efficiency loss from undertreatment by B firms is the loss in consumer welfare minus the cost savings from B performing a simpler treatment, times the probability of the undertreatment state, or $ph(v - \Delta c_B)$. The efficiency loss from overtreatment by S firms is just the extra resources wasted by S firms times the probability of the overtreatment state, or $(1-p)(1-h)\Delta c_S$. Thus overtreatment is more efficient than undertreatment when $\frac{ph}{(1-p)(1-h)} > \frac{\Delta c_S}{v-\Delta c_B}$. Thus, in efficiency terms, society prefers overtreatment more when p,h, and v are larger, and when Δc_S and Δc_B are smaller. Given the above comparative statics, this means that a mistreating monopolist is likely to choose the more efficient direction of mistreatment. This is not a coincidence: more efficient mistreatment allows higher prices.

Note that in theory, all of these implications are testable empirically. However, these first results depend on many assumptions, some of which I now relax. The assumptions I relax are that (1) the expert rather than the consumer offers the contract and there is only one expert, which together can be thought of as an absence of competition, (2) the consumer cannot decline treatment post-diagnosis, and (3) cost function variation is exogenous. Throughout the discussion I maintain the assumption of negative correlation in cost functions.

III.D.4 Competition: Consumer Offers Contract

Suppose the consumer, rather than the expert, can offer price contracts to experts. There are two cases, a market with a single expert with uncertain cost type, and a market with at least one expert of both cost types. The former case is akin to a regulated monopoly problem, the second case more akin to competition.

First, I should point out the separate roles of multiple treatments and asymmetric information in creating problems for the credence good market. Consider a world with asymmetric information about cost types but only one treatment option. This is a standard problem in procurement. Let costs equal c_S or c_B for this treatment, where $c_S < c_B$. In the monopoly case, the consumer will set $q = c_S$ and risk foregoing trade with probability p if $v < c_S + \frac{c_B}{p}$, and otherwise will set $q = c_B$. (Note that this is about information rents, not adverse selection.) We will see that adding the second treatment with the credence good structure will yield similar cases in which the consumer risks foregoing trade, but also creates the different phenomenon of mistreatment, which either adds to the efficiency costs of foregone trade, or creates inefficiency even when trade is guaranteed, and is always present to some degree. It is possible to decompose the efficiency costs into the costs of foregone trade and the costs of mistreatment.

In the competitive case with only one treatment, the consumer will offer the price that equals the cost of the lower-cost provider, trade will always prevail, and higher-cost providers will be driven out of the market, yielding the efficient outcome. Now adding a second treatment with the credence good structure does relatively more damage: it both raises the possibility of foregone trade, and guarantees mistreatment. In the competitive case, all of these costs are due to the multiple-treatment credence good structure; neither mistreatment nor foregone trade would be present in

the simpler one-good model of competition with asymmetric information about firm costs.

In the case of multiple treatment options characterized by the credence good structure, does empowering consumers to offer contracts yield efficiency, i.e., eliminate foregone trade and mistreatment? The next proposition shows that it does not, under either monopoly or competition.

Proposition 3 *The consumer cannot offer any contracts that induce honest treatment from both types of expert.*

Proof. Consider any $q_S \in R2$ that S accepts. Note that B earns profit $q_{LS} - c_{LB}$ at q_S, and in $R4$, B's profits on both goods are equal; hence at q_B, B's profit on both goods must at least equal $q_{LS} - c_{LB}$ if he is to accept q_B over q_S. Hence B will only prefer $q_B \in R4$ to q_S if $q_{LB} > q_{LS}$. But by definition of $R4$, this implies $q_{HB} > q_{HS}$. This means S will prefer q_B to q_S, for the same reasons that tempted B to prefer q_S over q_B. To keep S happy, we must raise q_{HS}, which requires raising q_{LS}, which starts the process again. Thus there is no pair of contracts that both firms will accept and that induces honesty from both firms. QED. ■

Proposition 3 is pessimistic about credence good markets: giving the consumer full control over prices does not bring about efficiency. Prices are too blunt an instrument for the consumer to overcome her informational disadvantage regarding both what she needs and who she's dealing with. The following two propositions describe equilibria arising when consumers set prices. The first proposition describes the set of contracts the consumer might offer in the monopoly case. Which contract arises will depend on market parameters, as described in the proof in the Appendix.

Proposition 4 *The consumer will offer a single price vector. This price vector will induce one of the following four outcomes: (1) pooling with honest treatment by B firms and overtreatment by S firms, (2) pooling with honest treatment by S firms and undertreatment by B firms, (3) pure overtreatment provided at cost only by S firms, or (4) undertreatment provided at cost by B firms and overtreatment provided at cost by S firms. Which numbered outcome prevails depends on which*

of the following corresponding numbered expressions is largest:

$$\begin{aligned}
&(1) \quad (1-ph)\,v - [1-h\,(1-p)]\,c_{LS} - h\,(1-p)\,c_{HS} \\
&(2) \qquad v - [1-p\,(1-h)]\,c_{HB} - p\,(1-h)\,c_{LB} \\
&(3) \qquad (1-p)\,(v-c_{HS}) \\
&(4) \qquad (1-ph)\,v - pc_{LB} - (1-p)\,c_{HS}.
\end{aligned}$$

Proof. Follows from calculations of consumer utility, which yield these four contracts as the set of non-dominated contracts. See Appendix for details. ■

The next proposition describes the analogous set of contracts that arise when consumers offer contracts to a population of experts including at least one expert of each type. The results change because now consumers can offer contracts that only lower-cost providers accept without risking foregone trade.

Proposition 5 *The consumer will offer a single price vector. This price vector will induce one of the following four outcomes: (1) pooling with honest treatment by B firms and overtreatment by S firms, (2) pooling with honest treatment by S firms and undertreatment by B firms, (3) pure overtreatment provided at cost only by S firms, or (4) pure undertreatment provided at cost only by B firms. Which numbered outcome prevails depends on which of the following corresponding numbered expressions is largest:*

$$\begin{aligned}
&(1) \quad (1-ph)\,v - [1-h\,(1-p)]\,c_{LS} - h\,(1-p)\,c_{HS} \\
&(2) \qquad v - [1-p\,(1-h)]\,c_{HB} - p\,(1-h)\,c_{LB} \\
&(3) \qquad v - c_{HS} \\
&(4) \qquad (1-h)\,v - c_{LB}.
\end{aligned}$$

Proof. Follows from calculations of consumer utility, which yield these four contracts as the set of non-dominated contracts. See Appendix for details. ■

Propositions 4 and 5 show that giving the consumer all the power, which is closely related to competition, can yield a greater variety of outcomes than giving the expert all the power, but have

no clear effect on efficiency. More outcomes are possible when the consumer sets prices because now prices depend on the actual cost *levels* faced by firms, not just the consumer's WTP and the expert's cost *wedges*. The comparative statics from the firm-offers-contract case more or less continue to hold, but letting the consumer set prices can have strange effects. For example, there are cases in which the firm sets prices to yield overtreatment, while consumers set prices to yield undertreatment[94].

One feature of these equilibria is that in some cases experts fully reveal their types to consumers in the act of accepting contracts. This may seen unrealistic, in that a consumer who learned an expert's true type in the process of trade would update her beliefs and insist on a contract that set incentives correctly. There would even be surplus to share motivating both parties to change the contract. However, this raises the possibility that a firm might accept a contract it didn't actually want in order to induce renegotiation. I have not explored this more complicated dynamic game in the consumer-offers-contract case. However, the next section shows that this is also a problem when the expert offers the contract, and that eliminating this feature does not improve efficiency in that case.

III.D.5 Price Ceilings (Weak Commitment)

When the monopolist expert offers the contract, the pooling equilibria characterized in Proposition 2 involve $q_H > v$. This implies that the contract as written is time-inconsistent: the consumer wishes to renege on the contract when he receives the diagnosis H.

A time-consistent mechanism to implement this pricing scheme could simply set the diagnosis cost $d = q_L$ and then reparameterize prices to $q'_L = 0$, $q'_H = q_H - q_L$. Such a contract is time-consistent from the consumer's perspective and leaves incentives unchanged for the expert. This contract seems plausible: experts often charge large diagnostic fees.

Nevertheless, we may want to relax the commitment assumption in the current framework without allowing an endogenous diagnostic price. If the consumer can renege on the contract after diagnosis, we must add the constraints that $q_L \leq v$ and $q_H \leq v$. How does this affect the analysis?

The obvious thing to do is to charge $q_H = v$, and then raise q_L to compensate. But this is not

[94] The following case does the trick when consumers set prices for a monopolist expert: $v = 15$, $p = 0.5$, $h = 0.5$, $c_{LB} = 4$, $c_{HB} = 10$, $c_{LS} = 4$, $c_{HS} = 5$.

consistent with honesty. Incentive-compatibility requires a proportional reduction in both prices, lowering profits that experts achieve under honesty. Now no expert profits from honesty. The best an expert of either type can do under honest treatment is to charge $q_H = v$, earning the same markup on all treatments sold. This yields the same profit as optimal prices in the overtreatment region of the price space, and may yield more or less profits than prices inducing undertreatment.

Suppose S firms select honest prices $q_H = v$, $q_L = v - \Delta c_S$. Then B will pool on this price since $v - \Delta c_S - c_{LB} > v - c_{HB} \iff \Delta c_B > \Delta c_S$. This reduces the prices that S can charge while maintaining honesty, because consumers anticipate B's pooling and lower their estimated surplus. But S will not tolerate lower prices, because S was already indifferent between honesty and overtreatment. Therefore S will never price honestly. What about B? If B prices honestly, then S may or may not pool–it makes no difference to S or B. Therefore B might price honestly, but S will never price honestly.

The key comparison is whether experts prefer overtreatment or undertreatment. For S, this means $v - c_{HS} \gtrless (1-h)\, v - c_{LS} \iff hv \gtrless \Delta c_S$. Similarly for B, it means $hv \gtrless \Delta c_B$. Three cases emerge:

Case 1: $hv < \Delta c_S < \Delta c_B$: S and B both undertreat.

Case 2: $\Delta c_S < hv < \Delta c_B$: S overtreats, B undertreats.

Case 3: $\Delta c_S < \Delta c_B < hv$: S overtreats, B is indifferent between honesty and overtreatment.

Note that the comparative statics that held in the monopoly case once again carry over.

Therefore the ceiling on prices, which can loosely be interpreted as a relaxation of the commitment assumption, does not eliminate mistreatment, and plausibly increases it. A full relaxation of the commitment assumption would involve "second-opinions," which are beyond the scope of this paper.

III.D.6 Endogenous Cost Functions

In reality, experts have some control over their cost functions. For example, they can choose among available technologies and ownership structures. As an extreme case, suppose experts have perfect and costless control over their cost function. In the model this corresponds to a choice between S and B cost functions. This also does not solve the credence good problem, and might make things

worse.

First consider the monopoly case. To sustain honest treatment by the S expert, we require that $h\left(q_L + \Delta c_S - c_{HS}\right) + (1-h)\left(q_L - c_{LS}\right) > q_L - c_{LB}$, implying $c_{LS} < c_{LB}$, which is a contradiction. To sustain honest treatment by the B expert, we require $h\left(q_L + \Delta c_B - c_{HB}\right) + (1-h)\left(q_L - c_{LB}\right) > q_L + \Delta c_B - c_{HS}$, implying $c_{HS} - c_{LB} > \Delta c_B$, again a contradiction. Therefore a monopolistic expert with a choice between S and B cost functions will never provide honest treatment. Instead, he will overtreat as a B type if $v > \frac{c_{HS} - c_{LB}}{h}$, and undertreat as an S type otherwise. Again, this preserves all the comparative statics of the monopoly case with exogenous cost functions.

Now consider the monopolistic and competitive consumer-offers-contract cases, in which the contract most-favored by the consumer must prevail. At any price vector inducing honesty by a type of expert, that type of expert would earn higher profits by acting dishonestly under that price vector as the other type. Now the consumer makes the same choice that the monopolist makes: induce overtreatment by type B if $v > \frac{c_{HS} - c_{LB}}{h}$, and undertreatment from type S otherwise.

III.E Discussion

The above results depend on the assumption that costs for low and high treatments negatively correlate across experts. This assumption seems weak, as experts who are worse at all treatments should be driven out of the market.

The model still places very high informational demands on consumers, even after relaxing observability of costs. In reality, consumers do not have a good sense of h, the distribution of various problems, or p, the distribution of firm cost functions, especially once we abandon the simple two-by-two structure assumed here. Weakening these assumptions seems likely to make things worse, not better. In fact, it seems possible that consumers would not even attempt to play the game described here, instead abandoning all hope of monitoring experts with price signals in favor of pure reliance on non-price constraints. This is analogous to concluding, in Akerlof's lemons model, that consumers trying to buy a used car do not attempt to do inference about product quality from product price. Instead, they only buy from friends (ethics), or only buy from dealerships with good reviews (reputation), or insist on submitting the car to inspection by a trusted mechanic prior to purchase (second-opinions). This is another reason we must interpret the model's comparative statics about mistreatment as suggestive at best.

Anything that mutes price signals sent to consumers affects mistreatment. Insurance is one example. Suppose consumers pay copayment β of all full prices. They then perceive prices βq instead of q. This is equivalent to increasing v by a factor of $\frac{1}{\beta}$. Since higher v tends to increase overtreatment, so does a smaller copayment rate. This phenomenon is distinct from moral hazard and interacts with it in unknown ways.

As mentioned earlier, I have ignored professional ethics, reputation, and second opinions. Undoubtedly these constraints are at least as important as the traditional price and quantity mechanisms explored in the asymmetric information literature. However, all three of these non-price constraints face problems that suggest a great deal of room for mistreatment when the price mechanism breaks down. Mistreatment–unlike mischarging–does not require stark violations of professional ethics. It merely requires that experts partly defer to unconscious or semi-conscious biases when making decisions under uncertainty (Gawande 2009). Reputation works poorly when consumers interact rarely with sellers and have trouble assessing quality, two conditions that predominate in credence good markets. Second opinions also may work poorly in the credence good context. Many credence goods solve urgent problems; it is hard to shop around for experts when you are sick, or your car doesn't work, or termites are eating your house. Experts consulted for second-opinions are also biased toward disagreement with prior opinions, since this may induce the consumer to seek treatment from the second expert. Consistent with this worry, experts often ask if the consumer has already received a diagnosis, what the diagnosis was, and who gave it–nothing like the i.i.d. draws assumed in the economics literature on second-opinions. And many consumers may hesitate to question the recommendation of an authority figure in the first place. Therefore, there is no strong *a priori* reason to believe that these other constraints can solve the problems described here.

If mistreatment is a serious problem, it raises the question of an optimal policy response. While a formal analysis of optimal policy is beyond the scope of this paper, it is worth referencing three policy initiatives that make more sense given the above results. First, the market might not provide sufficient incentives to compress input price variation by adopting salary pay instead of volume-based pay (e.g., fee-for-service, sales commissions, or profit-sharing). Second, the market might not provide sufficient private incentives to develop technologies that separate diagnosis from treatment (Christianson 2008). Third, experts might not have sufficient private incentives to collect and

publish data on their own performance in a standardized format in order to reduce informational asymmetries. This might provide an additional reason to encourage electronic record-keeping in healthcare markets.

Some predictions of the model could be tested without measuring mistreatment. For example, one could measure the response of prices and profits of some firms to exogenous changes in cost functions of other firms in the market. One example of such a cost shock is regulations that prohibit doctors from owning their own laboratories. Testing other predictions requires measurement of mistreatment. This data could be obtained as follows for a particular credence good market. First, identify suppliers with neutral economic incentives for providing the cheap and expensive treatments in question. This sounds hard, but may be easy. Some experts are paid on salary and supplied with inputs by a central administrator. Others have large alternative sources of income such as rich spouses, or are subject to intense scrutiny by other experts or the public. Second, estimate models of treatment rates on this population of experts, accounting for observable variation in consumer attributes across experts. Third, impute "honest" treatment rates to experts with non-neutral incentives for providing the cheap versus expensive treatments, using data on these experts' clientele. Finally, define mistreatment as the difference between observed and imputed treatment.

III.F Conclusion

Expenditures on credence goods are a large subset of expenditures on all goods. In particular, healthcare is both the textbook example of a credence good, and may be the world's most important good–now or eventually (Hall and Jones 2007). Financial and legal services, including mortgage sales, are also important industries involving credence. There is widespread perception of mistreatment in all of these industries, reflected in media reports and academic research. Fears of mistreatment seem consistent with everyday experience.

Yet there remains no simple, robust explanation for widespread mistreatment. In this paper I have shown that a realistic and traditional assumption–that consumers do not perfectly observe expert cost functions–generates pervasive mistreatment in credence good markets. When cost functions are not observed, prices do not signal markups. Thus prices do not alert consumers to the existence and direction of expert biases. Mistreatment emerges under a variety of reasonable assumptions, including competition and endogenous technology-adoption. By focusing on prices

and quantities instead of non-price mechanisms, the analysis puts the credence good problem on the same footing as the traditional lemons problem, and thereby puts mistreatment on the same footing as adverse selection. The analysis also generates a host of testable implications.

Some broad implications are that (1) widespread mistreatment in credence good markets would be easy to explain in theory, (2) non-price mechanisms such as professional ethics, reputation, and second opinions that constrain expert mistreatment may play a crucial role in credence good markets, and (3) if these constraints do not eliminate mistreatment, then various government interventions may improve welfare.

Potential interventions in credence good markets include, among many others, subsidized compression of input-price variation (i.e., subsidies for paying experts on salary and purchasing inputs on their behalf), mandatory and standardized record-keeping by experts, and development of better diagnosis technologies to separate diagnosis from treatment. While this paper establishes that experts have clear incentives to mistreat, the actual extent of mistreatment and the optimal policy response both remain open questions.

III.G Appendix

Preliminaries Under pooling, the participation constraint (PC), prices, and profits in all five regions (see Figure 3.2) are as follows. Note that I solve for the highest prices that satisfy the definition of the price region (thereby embodying the incentive properties, i.e. overtreatment or undertreatment by firms S and B) and that leave the consumer with at least zero surplus.

These regions are labeled by their number and the letter "P" for pooling.

Region 1P

PC: $(1-h)\,v = q_L$

$q_L = (1-h)\,v$

$q_H < (1-h)\,v + \Delta c_S$

$\pi_S = (1-h)\,v - c_{LS}$

$\pi_B = (1-h)\,v - c_{LB}$

Region 2P

PC: $(1-ph)\,v = [1-h\,(1-p)]\,q_L + h\,(1-p)\,q_H$

$q_L = (1-ph)\,v - h\,(1-p)\,\Delta c_S$

$q_H = (1-ph)\,v + [1-h\,(1-p)]\,\Delta c_S$

$\pi_S = (1-ph)\,v + ph\Delta c_S - hc_{HS} - (1-h)\,c_{LS}$

$\pi_B = (1-ph)\,v - h\,(1-p)\,\Delta c_S - c_{LB}$

Region 3P

PC: $(1-ph)\,v = pq_L + (1-p)\,q_H$

Prices and profits in this region can vary and I do not list them here.

Region 4P

PC: $v = p\,(1-h)\,q_L + [ph+1-p]\,q_H$

$q_L = v - [1-p\,(1-h)]\,\Delta c_B$

$q_H = v + p\,(1-h)\,\Delta c_B$

$\pi_S = v + p\,(1-h)\,\Delta c_B - c_{HS}$

$\pi_B = v - (1-h)\,(1-p)\,\Delta c_B - hc_{HB} - (1-h)\,c_{LB}$

Region 5P

PC: $v = q_H$

$q_L < v - \Delta c_B$

$q_H = v$

$\pi_S = v - c_{HS}$

$\pi_B = v - c_{HB}.$

Under separation, the PC breaks into three regions for each type of firm: above, on, and below the incentive-compatible (IC), or equal markup, price line in (q_H, q_L) space. I label these three regions their number and the letter "S" (do not confuse this S with expert type S). If I refer to a region as just "Rx" instead of "RxP" or "RxS" then I am not specifying whether the price in this

region is a pooling or separating equilibrium. The prices and profits of each firm in each region under separation are as follows:

Region 1S $q_L = (1-h)\,v$

$q_H < (1-h)\,v + \Delta c_S$

$\pi_S = (1-h)\,v - c_{LS}$

$\pi_B = (1-h)\,v - c_{LB}$

Region 2S $q_{Lj} = v - h\Delta c_j$

$q_{Hj} = v + (1-h)\,\Delta c_j,\ j = S, B$

$\pi_S = v - hc_{HS} - (1-h)\,c_{LS}$

$\pi_B = v - hc_{HB} - (1-h)\,c_{LB}$

Region 3S $q_L < v - \Delta c_B$

$q_H = v$

$\pi_S = v - c_{HS}$

$\pi_B = v - c_{HB}.$

III.G.1 Proof of Proposition 2

The proof proceeds as follows. In each region, I ask if a deviation from the candidate overtreatment equilibrium in R4P to that region can be rendered unprofitable by some belief. I then check to make sure that this belief is consistent with the Intuitive Criterion, i.e., only puts weight on expert types for whom the deviation is not weakly dominated by the equilibrium. If the deviation is weakly dominated by the equilibrium for all types then the Intuitive Criterion imposes no restrictions.

Note that a price vector is a strategy. The only way beliefs affect the profit earned at that strategy is by inducing the consumer to accept or reject the offer. In each price region, for each belief, there are three subregions: inside the consumer's PC, on the PC, and past the PC. Prices past the PC are always dominated by the equilibrium because the consumer rejects the contract, yielding zero contracts. Prices inside the PC turn out not to matter. The key prices to check, as candidate deviations, in each price region and for each belief are those that lie on the consumer's PC.

Overtreatment Equilibrium

$R1$ **Deviations** S must not deviate to $R1$. Beliefs don't matter here: for all beliefs the consumer expects to be mistreated by both expert types.

$$\begin{aligned}
&\pi_S(R4P) > \pi_S(R1S) = \pi_S(R1P) \\
&\iff v - c_{HS} + p(1-h)\Delta c_B > (1-h)v - c_{LS} \\
&\iff \\
&\qquad v > \frac{1}{h}\left[\Delta c_S - p(1-h)\Delta c_B\right]. \qquad (35)
\end{aligned}$$

B must not deviate to $R1$.

$$\begin{aligned}
&\pi_B(R4P) > \pi_B(R1S) = \pi_B(R1P) \\
&\iff v - hc_{HB} - (1-h)c_{LB} - (1-h)(1-p)\Delta c_B > (1-h)v - c_{LB} \\
&\iff \\
&\qquad v > \Delta c_B\left[1 + \frac{(1-h)(1-p)}{h}\right]. \qquad (36a)
\end{aligned}$$

Note that the B condition implies the S condition, by the assumption that $\Delta c_B > \Delta c_S$.

We do not need to check any other prices in $R1$ because they are either rejected by consumers, or they are less tempting than prices on the PC, because they don't affect consumer beliefs or behavior, just extract less surplus. So here deviations to prices inside the PC only exist if there are deviations to prices on the PC.

$R5$ **Deviations** Beliefs don't matter here: for all beliefs the consumer expects to be mistreated by both expert types.

B does not deviate to $R5$ if:

$\pi_B(R4P) > \pi_B(R5)$

$\iff v - hc_{HB} - (1-h)c_{LB} - (1-h)(1-p)\Delta c_B > v - c_{HB}$

$\iff p > 0$, which is true by assumption.

S does not deviate to $R5$ if:

$\pi_S(R4P) > \pi_S(R5)$

$\iff v - c_{HS} + p(1-h)\Delta c_B > v - c_{HS}$

$\iff p(1-h)\Delta c_B > 0$, which is true by assumption.

Therefore all beliefs at all price vectors in $R5$ support the equilibrium.

$R2$ **Deviations** Check if $\mu = 1$ in $R2$ can support the equilibrium at the highest prices this belief can support. This is the belief least favorable to the deviation in this region, implying the most dishonesty. It is therefore the best candidate for a belief in this region to support the equilibrium.

D3. No deviations to $R2$ with beliefs $\mu = 1$.

B does not deviate to $R2$ with $\mu = 1$ if:

Same as condition (36a).

S does not deviate to $R2$ with $\mu = 1$ if:

$\pi_S(R4P) > \pi_S(R2S|\mu = 1)$

$\iff v - c_{HS} + p(1-h)\Delta c_B > (1-h)v - c_{LS}$

$\iff$ same as condition (35).

Therefore $\mu = 1$ at all $q \in R2$ can support the equilibrium under no additional restrictions.

Intuitive Criterion: To use this belief at all prices $q \in R2$, it must be the case that there is no price acceptable to consumers at which S wants to deviate given some belief and B does not want to deviate given any beliefs. This requires:

$\pi_B(R4P) > \pi_B(R2)$ (B does not deviate)

$\cap$ $\pi_S(R4P) < \pi_S(R2)$ (S deviates)

$\cap$ $q_H = q_L + \Delta c_S$ ($q \in R2$)

$\cap$ $(1-\mu)[hq_H + (1-h)q_L] + \mu q_L \leq v$ (PC)

$\cap$ $\mu \in [0,1]$.

B condition:

$\pi_B(R4P) > \pi_B(R2)$

$\iff v - hc_{HB} - (1-h)c_{LB} - (1-h)(1-p)\Delta c_B > q_L - c_{LB}$

$\iff q_L < v - \Delta c_B[h + (1-h)(1-p)]$.

S condition:

$\pi_S(R4P) < \pi_S(R2)$

$\iff v - c_{HS} + p(1-h)\Delta c_B < q_L - c_{LS}$

$\iff q_L > v - \Delta c_S + p(1-h)\Delta c_B$.

These two conditions imply a contradiction if $v - \Delta c_B[h + (1-h)(1-p)] < v - \Delta c_S + p(1-h)\Delta c_B \iff \Delta c_B > \Delta c_S$, which is true by assumption.

Therefore the belief $\mu = 1$ can sustain the equilibrium for all $q \in R2$.

$R3$ **Deviations** Suppose only one type would deviate to a particular price in $R3$ for at least some belief. Then at that price beliefs must put weight on that type only.

Suppose that belief puts all weight on S. Then the consumer expects overtreatment in $R3$.

Suppose that price is outside the PC for $R5$. Then the consumer rejects the contract and profit at this price is zero.

Suppose that price is inside the PC for $R5$. Then the $R5$ case, which checks the most profitable possible deviation that consumers will accept under under this belief, is sufficient to cover this price vector as well, since only q_H matters for S in $R3$.

Suppose that belief puts all weight on B. Then the consumer expects undertreatment in $R3$.

Suppose that price is outside the PC for $R1$. Then the consumer rejects the contract and profit at this price is zero.

Suppose that price is inside the PC for $R1$. Then the $R1$ case, which checks the most profitable possible deviation that consumers will accept under under this belief, is sufficient to cover this price vector as well, since only q_L matters for B in $R3$.

This establishes that price vectors in $R3$ that are weakly dominated by the equilibrium for at least one type require no additional restrictions not already imposed by prices in $R1$ and $R5$.

Suppose both types would deviate to a particular price in $R3$ for some beliefs. Then the Intuitive Criterion does not restrict beliefs.

Consider the conditions that S not deviate and B not deviate at such a price q. The conditions for this are:

$\pi_S(R3) < \pi_S(R4P)$ (S does not deviate)

$\cap\ \pi_B(R3) < \pi_B(R4P)$ (B does not deviate)]

$\cap\ q_L < q_H - \Delta c_S$ ($q \in R3$)

$\cap\ q_L > q_H - \Delta c_B$ ($q \in R3$)

$\cap\ \mu q_L + (1-\mu) q_H \leq (1-\mu h) v$ (PC)

$\cap\ \mu \in [0,1]$ (valid belief).

Get a restriction using the S condition:

$\pi_S(R3) < \pi_S(R4P)$

$\iff q_H - c_{HS} < v - c_{HS} + p(1-h)\Delta c_B$

$\iff v > q_H - p(1-h)\Delta c_B$.

Get a restriction using the B condition:

$\pi_B(R3) < \pi_B(R4P)$

$\iff q_L - c_{LB} < v - (1-h)(1-p)\Delta c_B - h c_{HB} - (1-h) c_{LB}$

$\iff v > q_L + [h + (1-h)(1-p)]\Delta c_B$.

Check if one is redundant:

$q_H - p(1-h)\Delta c_B < q_L + [h + (1-h)(1-p)]\Delta c_B$

$\iff q_L > q_H - \Delta c_B$, which is true in $R3$, meaning that

if B does not deviate, then S does not deviate

if S deviates, then B deviates.

So find a belief that makes B not deviate. B undertreats in $R3$. By construction, beliefs are not restricted at the price under consideration, so set $\mu = 1$. So the consumer assumes undertreatment. Then all prices in $R3$ that are inside the PC for $R1$ are rejected by the $R1$ condition, and all prices in $R3$ that are outside the PC for $R1$ are rejected by consumers. We know that this keeps B from deviating. And we know that if B does not deviate then S does not deviate.

The establishes that $\mu = 1$ supports the equilibrium for price vectors at which there exists some μ that makes S deviate and some μ that makes B deviate.

Undertreatment Equilibrium

$R1$ **Deviations** D1. S must not deviate to $R1$. Beliefs don't matter here.

$$\pi_S(R2P) > \pi_S(R1S) = \pi_S(R1P)$$
$$\iff (1-ph)v - hc_{HS} - (1-h)c_{LS} + ph\Delta c_S > (1-h)v - c_{LS}$$
$$\iff v > \Delta c_S, \text{ which is true by assumption.}$$

B must not deviate to $R1$.

$$\pi_B(R2P) > \pi_B(R1S) = \pi_B(R1P)$$
$$\iff (1-ph)v - c_{LB} - h(1-p)\Delta c_S > (1-h)v - c_{LB}$$
$$\iff v > \Delta c_S, \text{ which is true by assumption.}$$

$R5$ **Deviations** D2. No deviations to $R5$ (does not depend on beliefs).

B does not deviate to $R5$ if:

$$\pi_B(R2P) > \pi_B(R5)$$
$$\iff (1-ph)v - c_{LB} - h(1-p)\Delta c_S > v - c_{HB}$$
$$\iff$$
$$v < \frac{1}{ph}\left[\Delta c_B - h(1-p)\Delta c_S\right]. \tag{37}$$

S does not deviate to $R5$ if:

$$\pi_S(R2P|\mu = p) > \pi_S(R5)$$
$$\iff (1-ph)v - hc_{HS} - (1-h)c_{LS} + ph\Delta c_S > v - c_{HS}$$
$$\iff$$
$$v < \frac{1-h(1-p)}{ph}\Delta c_S. \tag{38}$$

Note that the S condition implies the B condition, by the assumption that $\Delta c_B > \Delta c_S$.

$R3$ **Deviations** Suppose only one type would deviate to a particular price in $R3$ for at least some belief. Then at that price beliefs must put weight on that type only.

Suppose that belief puts all weight on S. Then the consumer expects overtreatment in $R3$.

Suppose that price is outside the PC for $R5$. Then the consumer rejects the contract and profit at this price is zero.

Suppose that price is inside the PC for $R5$. Then the $R5$ case, which checks the most profitable possible deviation that consumers will accept under under this belief, is sufficient to cover this price vector as well, since only q_H matters for S in $R3$.

Suppose that belief puts all weight on B. Then the consumer expects undertreatment in $R3$.

Suppose that price is outside the PC for $R1$. Then the consumer rejects the contract and profit at this price is zero.

Suppose that price is inside the PC for $R1$. Then the $R1$ case, which checks the most profitable possible deviation that consumers will accept under under this belief, is sufficient to cover this price vector as well, since only q_L matters for B in $R3$.

This establishes that price vectors in $R3$ that are weakly dominated by the equilibrium for at least one type require no additional restrictions not already imposed by prices in $R1$ and $R5$.

Suppose both types would deviate to a particular price in $R3$ for some beliefs. Then the Intuitive Criterion does not restrict beliefs.

Consider the conditions that S not deviate and B not deviate at such a price q. The conditions for this are:

$\pi_S(R3) < \pi_S(R2P)$ (S does not deviate)

$\cap\ \pi_B(R3) < \pi_B(R2P)$ (B does not deviate)]

$\cap\ q_L < q_H - \Delta c_S$ ($q \in R3$)

$\cap\ q_L > q_H - \Delta c_B$ ($q \in R3$)

$\cap\ \mu q_L + (1-\mu) q_H \leq (1-\mu h) v$ (PC)

$\cap\ \mu \in [0,1]$ (valid belief).

Get a restriction using the S condition:

$\pi_S(R3) < \pi_S(R2P)$

$\iff q_H - c_{HS} < (1-ph) v - h c_{HS} - (1-h) c_{LS} + ph\Delta c_S$

$\iff (1-ph) v > q_H - [1 - h(1-p)] \Delta c_S.$

Get a restriction using the B condition:

$\pi_B(R3) < \pi_B(R2P)$

$\iff q_L - c_{LB} < (1-ph) v - c_{LB} - h(1-p) \Delta c_S$

$\iff (1-ph) v > q_L + h(1-p) \Delta c_S.$

Check if one is redundant:

$q_L + h(1-p)\Delta c_S < q_H - [1 - h(1-p)]\Delta c_S$

$\iff q_L < q_H - \Delta c_S$, which is true in $R3$, meaning that

if S does not deviate, then B does not deviate

if B deviates, then S deviates.

So find a belief that makes S not deviate. S overtreats in $R3$. By construction, beliefs are not restricted at the price under consideration, so set $\mu = 0$. So the consumer assumes overtreatment. Then all prices in $R3$ that are inside the PC for $R5$ are rejected by the $R5$ condition, and all prices in $R3$ that are outside the PC for $R5$ are rejected by consumers. We know that this keeps S from deviating. And we know that if S does not deviate then B does not deviate.

This establishes that $\mu = 0$ supports the equilibrium for price vectors at which there exists some μ that makes S deviate and some μ that makes B deviate.

$R4$ **Deviations** Check if $\mu = 0$ in $R4$ can support the equilibrium at the highest prices this belief can support. This is the belief least favorable to the deviation in this region, implying the most dishonesty. It is therefore the best candidate for a belief in this region to support the equilibrium.

B does not deviate to $R4$ with $\mu = 0$ if:

$\pi_B(R2P) > \pi_B(R4|\mu = 0)$

$\iff (1-ph)v - c_{LB} - h(1-p)\Delta c_S > v - c_{HB}$

$\iff$ same as (37), see condition above.

S does not deviate to $R4$ with $\mu = 0$ if:

$\pi_S(R2P) > \pi_S(R4|\mu = 0)$

$\iff (1-ph)v - hc_{HS} - (1-h)c_{LS} + ph\Delta c_S > v - c_{HS}$

$\iff$ same as (38), see condition above.

Therefore, the belief $\mu = 0$ at all $q \in R4$ supports the equilibrium under no additional assumptions.

Intuitive Criterion: To use this belief at all prices $q \in R4$, it must be the case that there is no price acceptable to consumers at which B wants to deviate given some belief and S does not want to deviate given any belief. Such a price would require:

$\pi_S(R2P) > \pi_S(R4|\mu = 1)$ (S does not deviate)

$\cap$ $\pi_B(R2P) < \pi_B(R4|\mu = 1)$ (B deviates)

$\cap$ $q_L = q_H - \Delta c_B$ ($q \in R4$)

$\cap$ $\mu[hq_H + (1-h)q_L] + (1-\mu)q_H \leq v$ (PC)

$\cap$ $\mu \in [0,1]$.

Implies:

(S not deviate) $(1-ph)v - hc_{HS} - (1-h)c_{LS} + ph\Delta c_S > q_H - c_{HS}$

$\iff$ $q_H < (1-ph)v + (1-h)\Delta c_S + ph\Delta c_S$.

(B deviates) $(1-ph)v - c_{LB} - h(1-p)\Delta c_S < q_H - c_{HB}$

$\iff$ $q_H > (1-ph)v + \Delta c_B - h(1-p)\Delta c_S$.

This set is empty if:

$(1-ph)v + (1-h)\Delta c_S + ph\Delta c_S < (1-ph)v + \Delta c_B - h(1-p)\Delta c_S$

$\iff$ $\Delta c_S < \Delta c_B$, which is true by assumption.

Therefore the belief $\mu = 0$ can sustain the equilibrium for all $q \in R4$.

This establishes the proposition.

III.G.2 Proof of Proposition 4

Recall that in this proposition I assume the market contains only a single expert, whose type is unknown to the consumer.

Define C_B as a contract offered by the consumer that B accepts (i.e., implies $\pi_B \geq 0$) in equilibrium, and likewise for C_S, and define a pair of contracts $C \equiv \{C_S, C_B\}$. This is a complete strategy for the consumer.

Now consider other pairs of contracts, one by one. There are 5*5=25 ordered price region pairs to consider, and there may be multiple options within some of these ordered price region pairs. I will show that the consumer only needs to consider contracts in a subset of these pairs.

Proposition 3 shows that there does not exist incentive-compatible C such that $C_B \in R4$ and $C_S \in R2$.

Now consider separating contract pairs (i.e., C such that $C_S \neq C_B$) in which only one expert type acts honestly.

First, consider any $C_B \in R4$, $C_S \notin R4$. We have to offer S a contract that he prefers to C_B, but which B does not prefer to C_B.

Suppose $C_S \in R1$. To keep S happy we must set $q_{LS} - c_{LS} \geq q_{HB} - c_{HS}$. Thus the best the consumer could do would be to set $q_{LS} = q_{HB} - \Delta c_S > q_{HB} - \Delta c_B = q_{LB}$. But this doesn't work, because now $q_{LS} > q_{LB}$, so B prefers C_S to C_B.

Suppose $C_S \in R3$. This cannot work. Keeping S happy requires $q_{HS} > q_{HB}$, but in $R3$ this implies $q_{LS} > q_{LB}$, which induces B to prefer C_S to C_B.

Suppose $C_S \in R4$, $C_S \neq C_B$. If $q_{HS} > q_{HB}$ then B prefers C_S. If $q_{HS} < q_{HB}$, then S prefers C_B. Therefore incentive-compatible contracts pricing both B and S in $R4$ require pooling.

Suppose $C_S \in R5$. We can set $q_{HS} = q_{HB}$, and $q_{LS} = 0$, and now the contracts are separating. But this contract is trivial for the consumer, who gets the same surplus he would get if he just set $\pi_B = 0$ and offered $C_B = C_S \in R4$.

Conclusion: of the contracts involving $C_B \in R4$, we only need to consider pooling contracts $C_B = C_S \in R4$.

Now let's check this for $C_S \in R2$, and $C_B \notin R2$. First suppose $q_B \in R1$. This requires $q_{LB} \geq q_{LS}$. This is possible if $q_{LB} = c_{LS}$ and q_B is arbitrarily small. But this yields the consumer the same surplus as the pooling equilibrium with $q_B = q_S \in R2$. What about $q_B \in R3$? This also fails: the only way to make B prefer such a contract over $C_S \in R2$ is to raise both prices, which makes S prefer B's contract. What about $q_B \in R5$? Same problem as in $R3$.

Conclusion: of the contracts involving $C_S \in R2$ or $C_B \in R2$, we only need to consider pooling contracts $C_S = C_B \in R2$.

Therefore the consumer cannot offer $q_S \in R2$, $q_B \notin R2$, nor can he offer $q_B \in R4, q_S \notin R4$.

The remaining options are $q_S = q_B \in R2$, $q_S = q_B \in R4$, or any prices exclusively in mistreatment regions.

Consider $C_S \in R1$.

Suppose $C_B \in R1$. Both types choose the higher q_L. Therefore can pool at higher c_L or only trade with B at lower c_L.

Suppose $C_B \in R3$. Keeping S happy requires $q_{LS} - c_{LS} \geq q_{HB} - c_{HS}$ or $q_{LS} \geq q_{HB} - \Delta c_S$, and keeping B happy requires $q_{LB} \geq q_{LS}$. These conditions imply $q_{LB} \geq q_{HB} - \Delta c_S$, implying $C_B \in \{R1, R2\}$, which is a contradiction.

Suppose $C_B \in R5$. Same argument as $C_B \in R3$.

Conclusion: the only feasible contracts involving $C_S \in R1$ obey $C_S = C_B \in R1$. There are two such contracts to consider: one that induces trade from both S and B, and one that only induces trade from B, the lower-cost provider, and foregoes trade with S.

Consider $C_S \in R2$.

See above.

Consider $C_S \in R3$.

Suppose $C_B \in R1$. Keeping S happy requires $q_{HS} - c_{HS} \geq q_{LB} - c_{LS}$ or $q_{HS} \geq q_{LB} - \Delta c_S$, and keeping B happy requires $q_{LB} \geq q_{LS}$. Together, these conditions imply $q_{HS} \geq q_{LS} - \Delta c_S$, meaning $C_S \in \{R1, R2\}$, a contradiction.

Suppose $C_B \in R3$. We need $q_{LB} \geq q_{LS}$ and $q_{HS} \geq q_{HB}$. This requires q_B to lie northwest of q_S in (q_H, q_L) space. Consider the price vector $q = (q_L, q_H) = (c_{LB}, c_{HS})$. This price vector is weakly incentive-compatible; offering B price vectors just to the west and S price vectors just to the south would be strictly incentive-compatible. In addition, these prices induce B to undertreat at cost and S to overtreat at cost. This is the best the consumer can do in $R3$. The question is whether these price vectors actually exist in $R3$. They do. To show this, I show that q is strictly inside $R3$, meaning there is room to the west and south in which to write strictly incentive-compatible contracts as proposed. Price q is strictly inside $R3$ if $q_L \in (q_H - \Delta c_B, q_H - \Delta c_S)$. This requires $c_{LB} \in (c_{HS} - \Delta c_B, c_{HS} - \Delta c_S)$, which is true by the assumption that $\Delta c_S < \Delta c_B$. Therefore, separating contracts inducing lowest-cost mistreatment from both types of providers with $C_S \in R3$ and $C_B \in R3$ are feasible and must be considered. This is a "specialization" equilibrium in which consumers go to low-cost butchers who always sell them the service they're good at, whether or not that's the right service to provide.

Note that the $R3$ nature of this price is not important. This is equivalent to separating contracts inducing trade from B in $R1$ and S in $R5$. The range of prices to consider that are never actually charged in these specialization equilibria run from the pooling contract in $R3$ (or barely separating to induce strict incentive-compatibility) to setting these prices to zero, which puts B in $R1$ and S in $R5$. And we don't have to consider the "inefficient specialization" contracts with B in $R5$ and S in $R1$, because they are not incentive-compatible.

Suppose $C_B \in R5$. Here S requires $q_{HS} \geq q_{HB}$ and B requires $q_{HB} - c_{HB} \geq q_{LS} - c_{LB}$ or

$q_{HB} \geq q_{LS}+\Delta c_B$, together implying $q_{HS} \geq q_{LS}+\Delta c_B$ or $q_{LS} \leq q_{HS}-\Delta c_B$, meaning $C_S \in \{R4, R5\}$, a contradiction.

Conclusion: For C_S in $R3$, the only feasible contract also has $C_B \in R3$, and the best feasible contract for the consumer induces overtreatment at cost and undertreatment at cost from the lowest-cost providers. And also my old comment below about this being the same as four other possible contracts is correct.

Consider $C_S \in R4$.

See above.

Consider $C_S \in R5$.

Suppose $C_B \in R1$. We don't need to think about this, because it's the same as $R3$. All contracts that induce at-cost OT by S and at-cost UT by B offer the same payoffs.

Suppose $C_B \in R3$. Same, don't need to consider this.

Suppose $C_B \in R5$. Fine. Can pool at higher c_H or only trade with S at lower c_H.

Conclusion: only have to consider two contracts, both in $R5$, inducing treatment from one or both types.

Note that I do not have to repeat the above calculations again for C_B in $R1, R3, R5$. They are already considered, because there is nothing about the above calculations that is "from S's perspective."

Now I introduce some notation. Let $U(Rx)$ refer to maximum utility attainable (i.e., value function) at a pooling contract in region x, let $U(Rx, Ry)$ refer to the maximum utility attainable at a separating contract with q_S in region x and q_B in region y. Note that $U(Rx, \emptyset)$ reflects a contract that pushes B out of the market, and is thus trivially separating.

The above discussion shows that the consumer only needs to consider seven types of contracts. These contracts are:

$(R1)$

$(\emptyset, R1)$

$(R2)$

$(R3)$

$(R4)$

$(R5)$

$(R5, \emptyset)$.

We can calculate the utility available in the best (for the consumer) available versions of all these feasible contracts and compare them to find the utility-maximizing contract.

I first calculate the utility available at pooling contracts in every region.

$U(R2) = p[h(0 - c_{LS}) + (1-h)(v - c_{LS})] + (1-p)[v - hc_{HS} - (1-h)c_{LS}]$

$= (1-ph)v - c_{LS} - h(1-p)\Delta c_S.$

$U(R4) = p[v - hc_{HB} - (1-h)c_{LB}] + (1-p)[v - c_{HB}]$

$= v - c_{HB} + p(1-h)\Delta c_B.$

$U(\emptyset, R1) = p[(1-h)v - c_{LB}].$

$U(R1) = (1-h)v - c_{LS}.$

$U(R5, \emptyset) = (1-p)(v - c_{HS}).$

$U(R5) = v - c_{HB}.$

$U(R3) = p[(1-h)v - c_{LB}] + (1-p)[v - c_{HS}]$

$= (1-ph)v - pc_{LB} - (1-p)c_{HS}.$

Now we check all pairwise comparisons of feasible contract-pairs. Note that we don't actually have to check all pairs; it's a knock-out tournament, and knock-outs are transitive.

$U(R4) > U(R2) \iff phv > c_{HB} - c_{LS} - p(1-h)\Delta c_B - h(1-p)\Delta c_S.$

$U(R2) > U(1) \iff v > \Delta c_S$, true by assumption.

$U(R4) > U(5) \iff p(1-h)\Delta c_B > 0$, true by assumption.

$U(R2) > U(\emptyset, R1) \iff (1-p)v > h(1-p)\Delta c_S + c_{LS} - pc_{LB}.$

$U(R2) > U(R5, \emptyset) \iff p(1-h)v > pc_{HS} - [1 - h(1-p)]\Delta c_S.$

$U(R4) > U(\emptyset, R1) \iff v > (1-p)c_{HB} + \frac{ph}{1-p(1-h)}c_{LB}.$

$U(R4) > U(R5, \emptyset) \iff pv > c_{HB} - (1-p)c_{HS} - p(1-h)\Delta c_B.$

$U(R2) > U(R3) \iff [1 - h(1-p)]\Delta c_S > p(c_{HS} - c_{LB}).$

$U(R4) > U(R3) \iff phv > (1-p)(c_{HB} - c_{HS}) + ph\Delta c_B.$

$U(R3) > U(\emptyset, R1) \iff v > c_{HS}$, true by assumption.

$U(R3) > U(R5, \emptyset) \iff (1-h)v > c_{LB}.$

$U(R5, \emptyset) > U(\emptyset, R1) \iff (1 - 2p + ph)v > (1-p)c_{HS} - pc_{LB}.$

Survivors:

$U(R4) = v - c_{HB} + p(1-h)\Delta c_B$

$U(R2) = (1-ph)v - c_{LS} - h(1-p)\Delta c_S$

$U(R5, \emptyset) = (1-p)(v - c_{HS})$

$U(R3) = (1-ph)v - pc_{LB} - (1-p)c_{HS}$

These results establish the proposition.

III.G.3 Proof of Proposition 5

Now we extend Proposition 4 to the case of a market with at least one expert of each type in the market. The key difference here is that the consumer faces no risk of foregone trade when offering prices that exclude some types from trade.

The set of feasible contracts is the same as in Proposition 5; that discussion did not rely on having only one expert in the market.

In $R2$, both types of expert find the contract profitable, so the consumer can randomly choose one and the chance of getting each type is the same as if there were only one available. Same in $R4$.

$$\begin{aligned} U(R2) &= p[h(0 - c_{LS}) + (1-h)(v - c_{LS})] + (1-p)[v - hc_{HS} - (1-h)c_{LS}] \\ &= (1-ph)v - c_{LS} - h(1-p)\Delta c_S. \end{aligned}$$

$$\begin{aligned} U(R4) &= p[v - hc_{HB} - (1-h)c_{LB}] + (1-p)[v - c_{HB}] \\ &= v - c_{HB} + p(1-h)\Delta c_B. \end{aligned}$$

In the efficient $R1$, now the consumer will get treated no matter what:

$U(\emptyset, R1) = (1-h)v - c_{LB}$.

There is now no reason to offer the all-inclusive $R1$.

$U(R1) = (1-h)v - c_{LS}$.

In the efficient $R5$, the consumer will again get treated no matter what:

$U(R5, \emptyset) = v - c_{HS}$.

There is now no reason to offer the all-inclusive $R5$.

$U(R5) = v - c_{HB}$.

Utility does not change in $R3$.

$$U(R3) = p[(1-h)v - c_{LB}] + (1-p)[v - c_{HS}]$$
$$= (1-ph)v - pc_{LB} - (1-p)c_{HS}.$$

Now redo all the pairwise comparisons, ignoring the all-inclusive over- and under-treatment equilibria that are now strictly dominated. Note that the inclusive $R1$ and $R5$ were already dominated anyway. So the key change is that now $U(\emptyset, R1)$ and $U(R5, \emptyset)$ might dominate more things.

$U(R4) > U(R2) \iff phv > c_{HB} - c_{LS} - p(1-h)\Delta c_B - h(1-p)\Delta c_S.$

$U(R2) > U(\emptyset, R1) \iff v > \Delta c_S + \frac{c_{LS} - c_{LB}}{h(1-p)}.$

$U(R2) > U(R5, \emptyset) \iff v < \frac{1-h(1-p)}{ph}\Delta c_S.$

$U(R4) > U(\emptyset, R1) \iff v > \frac{1-p(1-h)}{h}\Delta c_B.$ Note these conditions (corrected) now look a lot like the Prop 2 and 3 conditions.

$U(R4) > U(R5, \emptyset) \iff p(1-h)\Delta c_B > c_{HB} - c_{HS}.$

$U(R2) > U(R3) \iff [1-h(1-p)]\Delta c_S > p(c_{HS} - c_{LB}).$

$U(R4) > U(R3) \iff phv > (1-p)(c_{HB} - c_{HS}) + ph\Delta c_B.$

$U(R3) > U(\emptyset, R1) \iff v > \frac{c_{HS} - c_{LB}}{h}.$

$U(R3) > U(R5, \emptyset) \iff v < \frac{c_{HS} - c_{LB}}{h}.$ Therefore $R3$ is either dominated by $(\emptyset, R1)$ or $(R5, \emptyset)$. Why? Because in $R3$, you get undertreated at cost with probability p or overtreated at cost with probability $1-p$, and you must prefer one of those outcomes to the other, so then you just get that one all the time. It's not so simple with only one seller, because you can't get one or the other all the time. Calculations show that $U(R5, \emptyset) > U(\emptyset, R1) \iff v > \frac{c_{HS} - c_{LB}}{h}$, confirming this explanation.

New survivors:

$U(R4) = v - c_{HB} + p(1-h)\Delta c_B$

$U(R2) = (1-ph)v - c_{LS} - h(1-p)\Delta c_S$

$U(\emptyset, R1) = (1-h)v - c_{LB}$

$U(R5, \emptyset) = v - c_{HS}$

This establishes the proposition.

REFERENCES

Alger, Ingela, and François Salanié. 2003. "A Theory of Fraud and Over-Consumption in Experts Markets." Boston College. Mimeo.

Akerlof, George A. 1970. "The Market for 'Lemons': Quality, Uncertainty, and the Market Mechanism." *Quarterly Journal of Economics*, 84(3): 488–500.

Baicker, Kate and Amitabh Chandra. 2004. "Medicare Spending, the Physician Workforce, and Beneficiaries' Quality of Care." *Health Affairs Web Exclusive*: W184–97.

Baron, David P and Roger B. Myerson. 1982. "Regulating a Monopolist with Unknown Costs." *Econometrica*, 50(4): 911–930.

Brownlee, S. 2008. *Overtreated: Why Too Much Medicine is Making Us Sicker and Poorer*. New York, NY: Bloomsbury Books USA.

Cho, I.K and D.M. Kreps. 1987. "Signaling Games and Stable Equilibria," *Quarterly Journal of Economics*, 102(2): 179-221.

Christensen, Clayton M., Jerome H. Grossman, Jason Hwang. 2008. *The Innovator's Prescription: A Disruptive Solution for Health Care*. McGraw-Hill.

Darby, Michael R., and Edi Karni. 1973. "Free Competition and the Optimal Amount of Fraud." *Journal of Law and Economics*, 16(1): 67–88.

Drake, R., J. Skinner, H. Goldman. 2008. "Commentary on 'What Explains the Diffusion of Treatments for Mental Illness?'," *American Journal of Psychiatry*, 165(11): 1385–1392.

Dulleck, Uwe and Rudolf Kerschbamer. 2006. "On Doctors, Mechanics, and Computer Specialists: the Economics of Credence Goods." *Journal of Economic Literature*, 44(1), pp. 5-42.

Ely, Jeffrey and Juuso Valimaki. 2003. "Bad Reputation," *Quarterly Journal of Economics*, 118(3): 785–814.

Emons, Winand. 1997. "Credence Goods and Fraudulent Experts." *RAND Journal of Economics*, 28(1): 107–19.

–2001. "Credence Goods Monopolists." *International Journal of Industrial Organization*, 19(3-4): 375–89.

Farrell, J. 1993. "Meaning and credibility in cheap talk games," *Games and Economic Behavior*,

5: 514-531.

Fisher, Elliot S., J.P. Bynum, and Jonathan S. Skinner. 2009. "Slowing the Growth of Health Care Costs – Lessons from Regional Variation," *New England Journal of Medicine*, 360(9): 849–852.

Fisher, Elliot S., David E. Wennberg, TA Stukel, DJ Gottlieb. 2004. Variations in the longitudinal efficiency of academic medical centers. Health Aff (Millwood); Suppl Web Exclusive: VAR 19-32.

Fisher Elliot S., David E. Wennberg, Stukel TA, DJ Gottlieb, FL Lucas, EL Pinder. 2003. "The Implications of Regional Variations in Medicare Spending. Part 1: the Content, Quality, and Accessibility of Care." *Annals of Internal Medicine*, 138: 273–287.

–2003. "The Implications of Regional Variations in Medicare Spending. Part 2: Health Outcomes and Satisfaction With Care." *Annals of Internal Medicine*, 138: 288–298.

Fong, Yuk-Fai. 2005. "When Do Experts Cheat and Whom Do They Target?" *Rand Journal of Economics*, 36(1): 113–30.

Fuchs, Victor R. 1978. "The Supply of Surgeons and the Demand for Operations." *Journal of Human Resources*, 13: 35–56.

Garber, A. and J. Skinner. 2008. "Is American Health Care Uniquely Inefficient?" *Journal of Economic Perspectives,* 22(4): 27–50.

Gawande, Atul. 2009. "The Cost Conundrum," *The New Yorker*, June 1. http://www.newyorker.com/reporting/

Gruber,Jon, John Kim, and Dina Mayzlin. 1999. "Physician Fees and Procedure Intensity: The Case of Cesarean Delivery." *Journal of Health Economics*, 18(4): 473–490.

Hall, Robert and Charles Jones. 2007. "The Value of Life and the Rise in Health Spending." *Quarterly Journal of Economics*, 122: 39-72.

Hughes, David, and Brian Yule. 1992. "The Effect of Per-Item Fees on the Behaviour of General Practitioners." *Journal of Health Economics*, 11(4): 413–37.

Jauhaur, S. 2009. "Referral Systems Turn Patients Into Commodities." *New York Times*, 5/26.

Kessler, Daniel and Mark McClellan. 1996. "Do Doctors Practice Defensive Medicine?" *Quarterly Journal of Economics*, 111(2): 353-90.

Mitford, J. 1998. *The American Way of Death Revisited.* Knopf.

Nelson, Phillip. 1970. "Information and Consumer Behavior." *Journal of Political Economy*,

78(2): 311–29.

Pitchik, Carolyn, and Andrew Schotter. 1987. "Honesty in a Model of Strategic Information Transmission." *American Economic Review*, 77(5): 1032–36.

– 1988. "Honesty in a Model of Strategic Information Transmission: Correction." *American Economic Review*, 78(5): 1164.

– 1993. "Information Transmission in Regulated Markets." *Canadian Journal of Economics*, 26(4): 815–29.

Richardson, Hugh. 1999. "The Credence Good Problem and the Organization of Health Care Markets." Texas A&M University. Mimeo.

Sirovich, Brenda, Elliot S. Fisher. 2006. "Regional Variations in Health Care Intensity and Physicians' Perceptions of Care Quality." *Annals of Internal Medicine*, 144(9): 641-649.

Sirovich, Brenda, Patricia M. Gallagher, David E. Wennberg, Elliot S. Fisher. 2008. "Discretionary Decision Making By Primary Care Physicians And The Cost Of U.S. Health Care," *Health Affairs*, 27(3): 813–823.

Skinner, Jonathan S., Douglas O. Staiger, Elliot S. Fisher. 2006. "Is Technological Change In Medicine Always Worth It? The Case Of Acute Myocardial Infarction." *Health Affairs,* 25(2): w34–w47.

Skinner, Jonathan S. 2009. "The Implications of Variations in Medicare Spending for Health Care Reform," Invited Testimony, Committee on Energy and Commerce, U.S. House of Representatives.

Sülzle, Kai, and Achim Wambach. 2005. "Insurance in a Market for Credence Goods." *Journal of Risk and Insurance*, 72(1): 159–76.

Taylor, Curtis R. 1995. "The Economics of Breakdowns, Checkups, and Cures." *Journal of Political Economy*, 103(1): 53–74.

Yasaitis, L., Elliot S. Fisher, Jonathan S. Skinner, and Amitabh Chandra, 2009. "Hospital Quality And Intensity of Spending: Is There An Association? Hospitals' Performance on Quality of Care is Not Associated With the Intensity of Their Spending," *Health Affairs*, 28(4), w566-w572.

Wolinsky, Asher. 1993. "Competition in a Market for Informed Experts' Services." *RAND Journal of Economics*, 24(3): 380–98.

– 1995. "Competition in Markets for Credence Goods." *Journal of Institutional and Theoretical Economics*, 151(1): 117–31.